Semantics and Similarity for Crowdsourced Geospatial Data

by

Andrea Ballatore

A Thesis submitted to
University College Dublin,
National University of Ireland
for the degree of Ph. D.
in the College of Science

January 2013

School of Computer Science and Informatics
John Dunnion (Head of School)
Under the supervision of
Michela Bertolotto, Ph. D.

In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it.

Jorge Luis Borges
On exactitude in science, 1946

Consider for example the proceedings that we call 'games.' I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all? ... And the result of this examination is: we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail.

Ludwig Wittgenstein
Philosophical Investigations, 1953

All our reasonings concerning matter of fact are founded on a species of Analogy, which leads us to expect from any cause the same events, which we have observed to result from similar causes. Where the causes are entirely similar, the analogy is perfect, and the inference, drawn from it, is regarded as certain and conclusive. ... But where the objects have not so exact a similarity, the analogy is less perfect, and the inference is less conclusive; though still it has some force, in proportion to the degree of similarity and resemblance.

David Hume
An Enquiry Concerning Human Understanding, 1748

To my mother, and to the memory of my father

ABSTRACT

Over the past decade, Volunteered Geographic Information (VGI) has blurred the boundaries between consumers and producers of geographic information. OpenStreetMap (OSM), the leading VGI project, aims at creating a map of the entire planet. The map semantics is negotiated by contributors through a Web 2.0 open tagging model, resulting in a fragmented and often ambiguous conceptualisation, which constitutes a barrier to the effective usage of OSM as source of geographic knowledge. This thesis develops a detailed and explicit conceptualisation for OSM, supporting the determination of the semantic relevance of concepts in geographic information retrieval, data mining, information integration, map personalisation, and user profiling. First, we design and extract a novel resource, the OSM Semantic Network, from the wiki website on which contributors negotiate the shared meaning of OSM geographic concepts. Second, we devise the Network-Lexical Similarity Measure (NetLexSiM), a hybrid approach to computing semantic similarity between concepts, through the linear combination of two aspects: network similarity and lexical similarity, drawing from graph-theoretical and natural language processing techniques. A new dataset, the Geo Relatedness and Similarity Dataset (GeReSiD), is created and used as a ground truth to demonstrate the high cognitive plausibility of the approach.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis would not have been possible without the active contribution of several people. First and foremost, I wish to thank my research supervisor, Dr. Michela Bertolotto, for her availability and unfailing support throughout my PhD, and the numerous opportunities that she created for me, including publications, conference attendances, and teaching. I thank Prof. David C. Wilson (Department of Software and Information Systems, University of North Carolina, Charlotte) for his weekly feedback on my work from overseas, and his very many insightful comments and suggestions.

Special thanks must be given to Science Foundation Ireland, for the generous financial support that allowed me to conduct my research free from the burden of poverty. I thank Prof. Kazutoshi Sumiya (School of Human Science and Environment, University of Hyogo) for accepting to be my external examiner despite the considerable geographic distance between Ireland and Japan. Furthermore, I thank Prof. Leslie Daly (UCD School of Public Health, Physiotherapy & Population Science) for his insights on the statistical aspects of my work.

Thanks go to my colleagues at the UCD School of Computer Science and Informatics, Gavin McArdle, and Ali Tahir, with whom I shared many meetings, meals, coffee breaks, and aimless research-centred chats. A mention goes to Alejandra López Fernández, who patiently followed and corrected my musings about natural language processing.

All of this would not have happened if my family in Italy, especially my mother Carla and my uncle Pierfranco, had not insisted that I study when my young mind was attracted to alternative plans. I am grateful to them for their unconditional support of my academic pursuits, even if the finer details of my research are not always clear to them. I hope that this thesis will somewhat justify in their eyes my prolonged stay in the rainy lands of Northern Europe. I also wish to thank my recently-acquired Irish family, Anne, Leslie, and Erika, for providing countless stimulating dinner table conversations – and nice food.

Last but certainly not least, I thank my girlfriend Selena for her prompt help with the vagaries of the English language, for her constant much needed support, and for making everything better.

LIST OF PUBLICATIONS

Journal papers

A. Ballatore, G. McArdle, A. Tahir & M. Bertolotto, *A Comparison of Open Source Geospatial Technologies for Web Mapping*, International Journal of Web Engineering and Technology. Inderscience, 6(4), pp. 354-374, 2011.

A. Ballatore, D.C. Wilson & M. Bertolotto, *Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap*, Knowledge and Information Systems, pp. 1-21, 2012.

A. Ballatore, D.C. Wilson & M. Bertolotto, *A Lexical Semantic Similarity Measure for Volunteered Geographic Concepts*, International Journal of Geographical Information Science (IJGIS), IN PRESS, 2013.

Conference proceedings

A. Ballatore, G. McArdle, C. Kelly & M. Bertolotto, RecoMap: An Interactive and Adaptive Map-Based Recommender. In *Proceedings of the 2010 ACM Symposium on Applied Computing* (SAC'10), pp. 887-891. Sierre, Switzerland, March 22-26, 2010.

A. Tahir, G. McArdle, A. Ballatore & M. Bertolotto, Collaborative Filtering - A Group Profiling Algorithm for Personalisation in a Spatial Recommender System. In *Proceedings of Geoinformatik 2010*, pp. 44-50. Kiel, Germany, March 17-19, 2010.

G. McArdle, A. Ballatore, A. Tahir & M. Bertolotto, An Open-Source Web Architecture for Adaptive Location Based Services, in *The 14th International Symposium on Spatial Data Handling* (SDH), vol. 38(2), pp. 296-301. Hong Kong, May 26-28, 2010.

A. Ballatore & M. Bertolotto, Semantically Enriching VGI in Support of Implicit Feedback Analysis, in *The 10th International Symposium on Web & Wireless Geographical Information Systems* (W2GIS), Springer, Lecture Notes in Computer Science, vol. 6574, pp. 78-93. Kyoto, Japan, March 3-4, 2011.

A. Ballatore, D.C. Wilson & M. Bertolotto, A Holistic Semantic Similarity Measure for Viewports in Interactive Maps, in *The 11th International Symposium on Web & Wireless Geographical Information Systems* (W2GIS), Springer, Lecture Notes in Computer Science, vol. 7236, pp. 151-166. Naples, Italy, April 12-14, 2012.

A. Ballatore, D.C. Wilson & M. Bertolotto, The Similarity Jury: Combining expert judgements on geographic concepts, in *The Proceedings of the 6th International Workshop on Semantic and Conceptual Issues in GIS (SeCoGIS 2012)*, Springer, Lecture Notes in Computer Science, vol. 7518, pp. 231-240. Florence, Italy, 2012.

A. Ballatore, M. Bertolotto & D.C. Wilson, Grounding Linked Open Data in WordNet: The Case of the OSM Semantic Network, in *The 12th International Symposium on Web & Wireless Geographical Information Systems* (W2GIS), Springer, Lecture Notes in Computer Science, vol. 7820, pp. 1-15. Banff, AB, 2013.

Book chapters

A. Ballatore, D.C. Wilson & M. Bertolotto, A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web, in *Quality Issues in the Management of Web Information*, Intelligent Systems Reference Library, vol. 50, Springer, pp. 93-120, 2013.

CHAPTER

ONE

INTRODUCTION

1.1 Overview

On January 31, 1998, former U.S. Vice-President Al Gore delivered a speech at the California Science Center, Los Angeles. He argued that rapidly improving digital technologies allow us to store, process and retrieve an unprecedented amount of information about the planet. The surface of the Earth is being monitored by an ever increasing number of sensors, capturing a wide variety of environmental and cultural phenomena [101]. To tap effectively this informational wealth, Gore envisaged the development of the 'Digital Earth,' an interactive, digital model of the whole Earth, centralising vast amounts of georeferenced data. Unlike other computer-centred utopian visions, as British geographer Michael Goodchild [98] pointed out, aspects of Gore's Digital Earth have materialised at a surprisingly rapid pace.

From 2001 onward, a nexus of online geographic services and tools has emerged, including NASA World Wind, CIA-funded Keyhole (later branded as Google Earth), Google Maps in 2005, later joined by Yahoo! Maps and Bing Maps, resulting in the so-called 'GeoWeb' [114]. The unique combination of the explosive growth of digital information, Web 2.0 and crowdsourcing offered fertile soil to create new forms of production and consumption of spatial information [155, 226, 126]. Although well-funded public and private institutions have given a key contribution to the GeoWeb, a novel figure has entered the production cycle in an intermediate position between the producers and the consumers, called 'produser' (producer/user) by Coleman et al. [51]. Similar to open source software contributors, these individuals devote unpaid labour and time to generate geographic information, relying on Web 2.0 collaborative tools.

Turner [291] named this resurgence in collaborative mapping 'neogeography', whilst Goodchild [96] has coined the term Volunteered Geographic Information (VGI). The most prominent case of VGI is without doubt the grassroots mapping project OpenStreetMap (OSM), in which experts and non-experts alike engage in the non-profit construction of a rich and increasingly complex vector dataset, covering the entire planet [113, 30]. Founded in 2004 in the UK, OSM has been collecting geographic data from a variety of sources, including existing public domain datasets and 'mapping parties' – the first was held at Isle of Wight in 2006. The size of the OSM vector map, the core asset

of the project, has been growing at an impressive rate – 150% over the last two years alone. On average, in 2011 over 96,000 kilometres of new roads were added to the dataset every week [223].

Indeed, it is difficult to deny the advantages that the rise of the GeoWeb has brought about for data consumers, both in its crowdsourced and expert-authored manifestations. Individual users and organisations have gained access to an unprecedented amount of geo-information repositories, opening up new avenues for scientific, commercial, and social mapping applications. The commercial boom of location-aware smartphones has driven the process further, resulting in what has been called 'ubiquitous cartography' [89]. This dramatic expansion of the online geographic landscape is inevitably creating new challenges for providers and consumers alike.

Whilst human beings are generally good at extracting meaning from text and maps, machines are considerably less intelligent. In particular, machines find it very difficult to interpret raw, unstructured data. In a seminal paper published in 2001 in the *Scientific American*, Berners-Lee et al. [31] imagined a Semantic Web in which data is expressed in machine-readable formats, allowing intelligent agents to "COMPREHEND semantic documents and data," with unthinkable benefits for information retrieval, knowledge management, classification, and so on. Over the past decade, this somewhat utopian vision has inspired wide-ranging efforts to develop and improve technologies for expressing complex data semantics [169, 42].

The urge for better data semantics soon attracted attention in the geographic domain, and Egenhofer [65] proposed a 'Semantic Geospatial Web' as a promising direction to improve the interaction with spatial data. While traditional geographic data is expert-authored, Web 2.0 and spatial crowdsourcing have added a growing amount of geo-data of varying quality and semantics [112]. As a result, the GeoWeb consists of a semantically fragmented and highly dynamic set of loosely interconnected contents, services, and interfaces, still far from the centralised 'geoportal' imagined by Al Gore [98].

The Semantic Web and Web 2.0 differ in many respects: if the Semantic Web promotes logic formalisms to enable automatic inferences, Web 2.0 relies on more informal semantics, such as collaborative tagging, linking, and social networking. However, as Greaves and Mika [104] pointed out, these two visions finally converge around the idea of 'socially shared meaning,' for which they provide different and yet complementary approaches. In a way, each approach can overcome some of its limitations via a further integration with the other [11]. In our work, we strengthen OSM's 'weak' approach to semantics, aiming at a machine-readable, rich description of crowdsourced concepts – the 'strong' approach to semantics [289].

The work presented in this thesis is located at this crossing between the Semantic Web and the Web 2.0, in the context of geographic information. Figure 1.1 depicts the key concepts and research areas that inform our work on OpenStreetMap (OSM) semantics, between these two main paradigms (Web 2.0 and Semantic Web) [104]. The Semantic Web, with its spatial subset, the Semantic Geospatial Web, relies on ontologies, and artificial intelligence (AI)

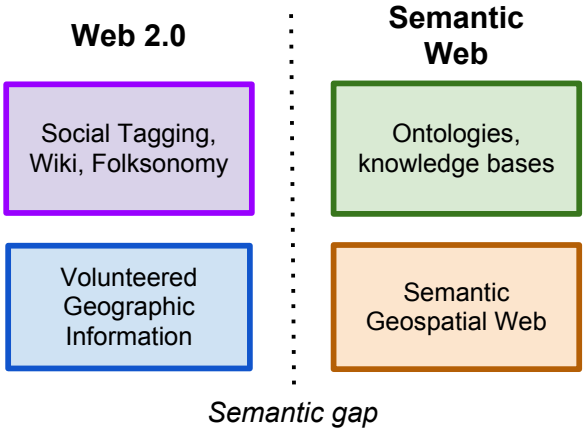

Figure 1.1: Semantic gap between Web 2.0 and the Semantic Web

formal languages ('strong' semantics). Web 2.0 and VGI, by contrast, collect data using loose, collaborative semantic models ('weak' semantics). Ideally, the informational value gathered by VGI could be utilised in both areas, in a virtuous interplay.

Our work aims at bridging the gap between the Semantic Geospatial Web technologies, and the crowdsourced, collaborative nature of OSM. The core original contribution of this thesis consists of an extension of the OSM semantics, the leading grassroots VGI project. First, we develop a novel semantic support tool, the OSM Semantic Network. Second, we devise a hybrid measure to compute the semantic similarity of OSM concepts, the Network-Lexical Similarity Measure (NetLexSiM). Finally, NetLexSiM is evaluated with a new human-generated dataset, the Geo Relatedness and Similarity Dataset (GeReSiD).

In the remainder of this chapter, we outline the core research problem in this thesis in Section 1.2, while Section 1.3 defines the purpose and the methodologies on which the research presented in this thesis is grounded. The nature and scope of our research contribution is discussed in Section 1.4. Finally, Section 1.5 describes the structure of this thesis.

1.2 Research problem

This section defines the central research problem which this thesis aims to tackle, namely, the computation of semantic similarity in crowdsourced spatial data. In a context of ever-expanding online geo-information, data semantics play a crucial role in enabling a wide range of applications. The more semantic content is *explicitly* modelled, the better an application can exploit the data.

The research problem was identified during preliminary work we conducted on map personalisation. In order to facilitate spatial tasks such as Geographic Information Retrieval (GIR), a promising and neglected direction

is that of map personalisation [187]. Personalisation characterises the post-industrial age, in which mass production has often been replaced by – real or perceived – personalised production [194]. Popular online service providers such as Google, Amazon and Facebook exploit personalisation techniques to deliver user-specific advertisements, search results, and recommendations [289]. In the geospatial domain, digital maps offer possibilities of personalisation unthinkable in paper-based traditional cartography. The individual user, instead of being a passive consumer of a pre-packaged map, should ideally become the very centre of the map generation process [242].

To put the user at the centre of an information system, a large body of theories and techniques have been devised around user modelling and profiling [157, 319]. An interactive mapping service can gather feedback from the user, build a user profile, and generate personalised maps based on the profile. Personalisation can be achieved through the use of *explicit feedback*, such as user preferences dialogs, in which the system asks the user to specify their interests and preferences. Although this approach can be very accurate, it requires explicit input, which can be time consuming, distorted by subjectivity and distract the user from their main task [312]. Feedback can also be collected by *implicit* techniques, aimed at ascertaining the user preferences from their interaction [305, 147]. The system monitors user interaction trying to determine their interests and builds a user profile. While having to deal with large amounts of noisy historical log data, this approach does not disrupt the current task of the user [187].

In our preliminary work on map personalisation, we noticed that the construction of user profiles for map personalisation involves remarkable geo-semantic challenges. In a typical session on web Geographic Information Systems (GIS), users explore geographic data by clicking on map features, submitting text searches, panning, and zooming in and out [123]. All this interaction happens at the *instance* level: users search for specific features, click on them, obtain routing directions, etc. For example, the system can profile a user by modelling the fact that she often searches for features described as 'university' or 'school' in the Dublin area. However, at this semantic level, the system stores explicit symbols (such as 'university' and 'school'), but cannot discover *implicit* underlying semantics, i.e. an interest in 'education.' This piece of information may be exploited to expand the user query (include objects described as 'college'), cluster the user with similar users (other people interested in education), recommend events, supporting countless applications.

This 'semantic gap' is typically solved through knowledge engineering. Ontologies, conceptual taxonomies and semantic networks can easily model the fact that 'university' and 'school' are both conceptually related to 'education.' However, in the context of VGI and spatial crowdsourcing, this level of explicit ontological modelling is rarely found. The leading VGI project, OSM, adopts a shallow semantic model, based on collaborative tagging. The lack of a centralised, formal ontology, is a crucial element to understand the project's growing popular success [223]. Not having a normative, top-down semantic framework allows contributors to negotiate a common ground in an open, col-

laborative process, on a dedicated wiki website.[1] More importantly, contributors' communities can construct their own semantics, accurately modelling local aspects of their geographic reality. As Goodchild [97] put it, "we are all experts in our own local communities" (p. 95). For example, the inhabitants of the Italian city of Turin are very familiar with the concept of *dehors* (roughly equivalent to 'pavement café' in English), which is likely to be unintelligible to most other Italians. The OSM loose semantics allows local concepts to be quickly added into the map.

This somewhat permissive, bottom-up approach to semantics has obvious drawbacks. If we are all experts in local geography, we are likely not to be so in any other place in the world, and thus local tags can be very difficult to interpret. Users often disagree on the tag semantics, and this sometimes results in 'tag wars' [211]. When it is necessary to align local data to the OSM conceptualisation, the process can be complex and frustrating, such as in the case of road classification.[2] More importantly from our perspective, users are not encouraged to provide a full ontological conceptualisation of the data semantics, but tend to adopt a minimalist approach, resulting in what can be called a 'folksonomy' [297]. The OSM concept 'school,' for example, do not contain any reference to the term 'student,' as human users are rightly expected to have an intuitive grasp of the relationship between students and schools.[3] As a result, nowhere in the OSM dataset there is a formal conceptualisation of concepts 'school' and 'university,' which could be exploited via Semantic Web technologies.

To enable an effective usage of OSM data, we deem that this semantic gap constitutes a key challenge to supporting map personalisation, GIR, and other applications. Filling this semantic gap in spatial data would help identify relevant information and would generate richer user profiles, tapping valuable crowdsourced data. To clarify the perspective from which we conducted our research, the following tenets summarise the research problem of this thesis:

- Geo-semantics is an important research area in Geographic Information Science (GIScience). Data semantics is considered crucial to utilise, integrate, index, and retrieve any geospatial data, especially when generated from diverse sources [161, 98].

- Crowdsourced spatial data poses a particular semantic challenge. The bottom-up nature of crowdsourced data tends to be incompatible with hierarchy, semantic accuracy, and formal ontologies, which are necessary to perform a wide range of complex tasks [32].

- The OpenStreetMap (OSM) semantics consists of a shallow model, based on open, collaborative tagging. Whilst this approach attracts contributors, it results in semantically ambiguous geographic data. This makes

[1] `http://wiki.openstreetmap.org` (All the URLs referenced in this thesis were accessed on November 4, 2012.)

[2] `http://wiki.openstreetmap.org/wiki/Highway:International_equivalence`

[3] `http://wiki.openstreetmap.org/wiki/Tag:amenity%3Dschool`

the usage of this data problematic in, among other applications, map personalisation, GIR, and information integration [113].

– OSM features are described using an open set of tags. The tags correspond to geographic concepts, such as 'lake' or 'university.' These concepts are not strictly formalised, but are loosely described by contributors on the OSM Wiki website. A machine-readable representation of the concepts would represent a useful semantic support tool for OSM users.

The bulk of the work presented in thesis aims at providing solutions to the open problem of crowdsourced semantics, through the development of a novel semantic network, and of a measure of semantic similarity for OSM tags. The next section will define the aims and the methodology we have followed in this thesis.

1.3 Aims and hypotheses

As discussed above, Volunteered Geographic Information (VGI) provides free-to-use large datasets, such as the OSM vector dataset, which can be used for a variety of purposes [96]. From a semantic viewpoint, crowdsourced spatial data tends to be *implicit* and ambiguous, and of highly varying quality [32]. By contrast, traditional knowledge engineering structures, such as semantic networks and ontologies, require *explicit* semantics and careful formalisation, and offer different degrees of inferential power [308, 86, 110].

In this context, our main goal consists of providing a general semantic support mechanism for spatial crowdsourced data, in the form of a bottom-up semantic network, which can be applied to information integration, ontology alignment, and data mining. Such semantic support for VGI may also benefit GIR, providing a rich dataset to assist the retrieval process of large sets of text documents [91, 192]. In order to reach this goal, we have focused on OSM, the most prominent VGI project.

First, we analysed in detail the semantic model underlying the OSM geographic dataset. The OSM vector dataset, the core asset of the project, is a large collection of points of interest (POI), polylines, and polygons, representing natural and man-made features. These objects are characterised by spatial content – geo-coordinates – and semantic content, i.e. what they actually represent (e.g. a park, a museum, a mountain, a post box, etc). The semantics of these geographic features is described through properties, having a key and a value, called 'tags' in the OSM jargon. For example, universities are marked as *amenity=university*, while residential roads as *highway=residential*.

The meaning of these tags is defined by crowdsourced descriptions on a dedicated website, the OSM Wiki website.[4] This website contains an *implicit* folksonomy, i.e. a bottom-up, semi-structured semantic network, whose pages correspond to concepts, and hyperlinks to semantic relations. Subsequently,

[4] `http://wiki.openstreetmap.org/wiki/Map_Features`

we reviewed the existing projects geared towards providing semantic support for OSM, such as LinkedGeoData [15] and OSMonto [50], to find out that, although valuable in many respects, such projects did not provide the extended semantic model we aim to develop. Therefore, to provide a semantic support tool for OSM users, we decided to extract a novel semantic network from the OSM Wiki website.

A semantic network representing OSM geographic concepts and their mutual relationships can be used as a support tool for a wide range of applications [308]. The computation of *semantic similarity* between two geographic concepts is an essential building block in our approach to personalisation. A computable approach to semantic similarity between geographic concepts can be exploited to uncover implicit interests (e.g. *universities* and *schools* are more similar than *universities* and *parks*, and should be clustered together).

Among semantic tasks that provide support for a wide range of applications, including map personalisation and GIR, we identified the computation of *semantic similarity* of geographic concepts. The ability to intuitively assess semantic similarity of concepts – e.g. 'school' and 'university' should be more similar than 'school' and 'mountain' – plays a fundamental role in human cognition and thinking [261, 135]. Therefore, we decided to tackle the issue of semantic similarity in the context of OSM, which can then provide semantic support for personalisation, data mining, GIR, and the numerous tasks that benefit from an effective measure of similarity. The semantic model of OSM appears simple and flexible, but does not allow a direct computation of semantic similarity. On the other hand, the semantic network contains lexical descriptions of concepts, which can be used to compute similarity by using natural language processing techniques. In summary, we set the following aims for this thesis, defining for each case relevant hypotheses to be tested:

Ontology Alignment. Explore alignment techniques between OSM and existing semantically rich knowledge bases such as DBpedia [208]. *Hypothesis*:

- OSM data can be linked to DBpedia through an unsupervised technique.

OSM Semantic Network. Develop a semantic network from the OSM Wiki website, the OSM Semantic Network, making it available online. Such a semantic network captures explicitly the relationships between geographic concepts, in a machine-readable semantic format. *Hypotheses*:

- The OSM Wiki website contains valuable semantic content about the OSM concepts, which can be mined and structured into a semantic network.
- The extracted semantic network can provide semantic support for a number of tasks, including the computation of semantic similarity.

Semantic similarity and relatedness. Compute semantic similarity and relatedness for OSM geographic concepts. Evaluate different semantic measures against human rankings, assessing their cognitive plausibility, i.e.

their ability to mimic human intelligent behaviour. Based on empirical evidence, we aim at providing pragmatic guidelines to researchers and users on the most effective techniques for semantic similarity. *Hypotheses*:

- Every concept is semantically related to all other concepts, but concepts used to describe the same domain of human experience are more related than other concepts.
- Tobler's first law of geography states that everything is related to everything else, but near things are more related than distant things [288]. Analogously, all concepts are semantically similar, but concepts defined with similar concepts are more similar than other concepts.

Network-Lexical Similarity Measure (NetLexSiM). Compute network-based and lexical semantic similarity for OSM geographic concepts; these two components are then combined in a hybrid measure. *Hypotheses*:

- Co-citation similarity measures applied on the OSM Semantic Network can reach high cognitive plausibility. Recursive measures perform better than non-recursive ones.
- The lexical definitions of OSM concepts allow the computation of a more plausible semantic similarity measure than the simple tags.
- Techniques for paraphrase detection, based on bag-of-words (BOW) techniques and WordNet similarity, can obtain high cognitive plausibility.
- The hybrid measure obtains higher cognitive plausibility than both network and lexical similarity.

Similarity jury. When multiple semantic similarity measures are available and the optimal measure is unknown, it is possible to combine them in a 'jury,' rather than relying on an individual measure. *Hypotheses*:

- A semantic similarity measure is comparable to a human expert expressing her opinion on a complex issue.
- A jury tends to perform better than the average individual expert, but the best individual in the jury often outperforms the jury as a whole.
- When the optimal measure is unknown, it is preferable to rely on a jury rather than on an arbitrary measure.

Viewport semantic similarity. From a GIR perspective, explore the possibility of treating viewports, and not single features, as semantic units in a information retrieval (IR) system. *Hypotheses*:

- A viewport has an implicit semantic content, which is not defined in any of its components.

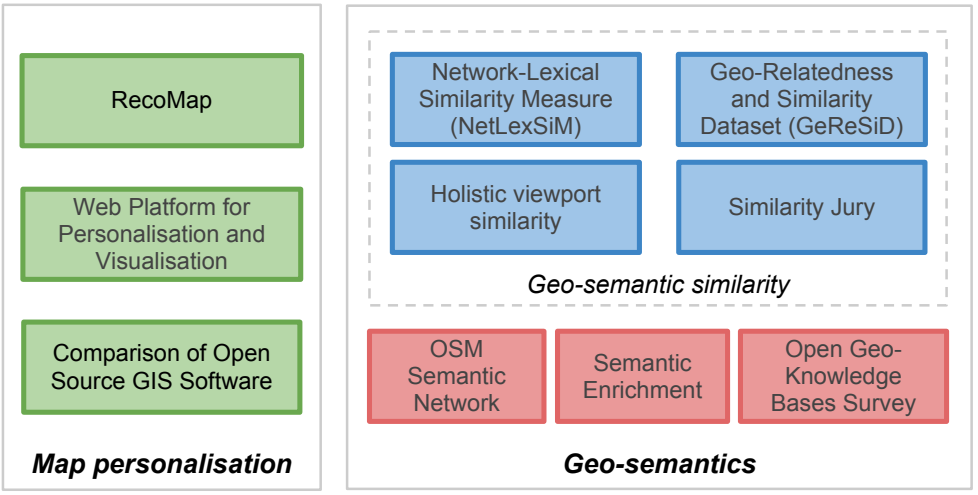

Figure 1.2: Outline of thesis contributions

– Such implicit semantics of map viewport can be captured using vector-based semantic descriptors.

These aims and hypotheses have defined a road map that led to the research presented in this thesis, developing a new semantic network for OSM, then utilised as a support tool to compute semantic similarity of concepts. After outlining the aims and roadmap we have followed in our work, the section will review the actual outcomes of our research efforts, detailing the aspects of the scientific contribution of the work contained in this thesis.

1.4 Thesis contribution

The original scientific contribution presented in this thesis consists of techniques for semantic enrichment and semantic similarity measures, in the framework of spatial crowdsourcing project OpenStreetMap (OSM). The contributions can be grouped in two phases: (1) a preliminary exploration of map personalisation, and (2) the study of semantic issues in crowdsourced spatial data. Figure 1.2 outlines the scientific contributions of this thesis.

First, we conducted research in the area of map personalisation, spatial recommender systems, and implicit feedback. This part of our work was conducted in collaboration with other research projects, in the framework of the Strategic Research in Advanced Geotechnologies (StratAG) cluster, funded by Science Foundation Ireland (SFI).[5] This work was essential to identify the semantic challenges then addressed in the core contribution of this thesis. The contributions that originated from this preliminary phase are the following:

– The *RecoMap* prototype: an approach to spatial recommendation based

[5] `http://www.stratag.ie`, `http://www.sfi.ie`

on implicit feedback. This system explores a novel technique to model the user interest in geographic features. The user interacts with a Web map, and her geographic interests are extracted dynamically from the interaction flow, and stored in a dynamic profile (Section 3.2).

– A technique to semantically enrich the OSM semantics exploiting existing open geo-knowledge bases. Starting from OSM features, this approach consists of retrieving the corresponding entities in DBpedia, a Semantic Web version of Wikipedia. This technique generates a richer geographic conceptualisation of the data contained in OSM, facilitating the exploitation of the data in information retrieval (IR) and data mining (Section 3.2).

– An online survey of free and open-source software (FOSS) Web mapping tools, including spatial DBMSs, Web GUI toolkits, and frameworks for Web applications, to poll their strengths and weaknesses. This survey provides direct feedback about a wide range of software packages that are commonly used in academia and industry (Section 4.3).

– A Web platform for map personalisation and visualisation: an open source architecture to implement and explore personalisation and implicit feedback techniques. This platform combines state-of-the-art technologies for Web mapping to provide a flexible environment to implement and deploy experimental personalisation and visualisation techniques (Section 4.4).

Subsequently, in the second phase, we focused on semantic aspects of OSM and Volunteered Geographic Information (VGI). To model user profiles for personalisation and implicit feedback, we identified an open problem. While objects can be considered as isolated semantic units, their relationships with other objects can increase the expressiveness and inferential power of a profiling technique. In other words, beyond the specific feature types, there are *implicit* semantic connections, whose knowledge can support a number of advanced spatial tasks. For example, if a user repeatedly queries a retrieval system to see schools and universities, her profile should model an interest in education. This piece of information can then be used to perform more accurate recommendations, search results, and group profiling. The exploration of this issue has resulted in a range of contributions in the area of geo-semantics, with particular focus on the OSM semantic model. In summary, the core contribution to knowledge in our thesis consists of the following points:

1. The OSM Semantic Network (Section 3.3): a crowdsourced semantic network representing OSM geographic concepts, extracted in a bottom-up process from the OSM Wiki website, released as Open Knowledge.[6] The dedicated crawler we have developed for this purpose is released under a GPL license.[7] This network is a machine-readable representation of the

[6] `http://github.com/ucd-spatial/OsmSemanticNetwork`
[7] `http://github.com/ucd-spatial/OsmWikiCrawler`

conceptualisation developed by OSM contributors, and has a wide range of applications.

2. Network-Lexical Similarity Measure (NetLexSiM): a novel approach to computing the semantic similarity of OSM concepts, i.e. 'tags,' in the OSM terminology, combining two components. The first component relies on existing network co-citation algorithms such as SimRank and P-Rank (Section 3.6). The second component, on the other hand, is a novel knowledge-base similarity measure, combining WordNet-based similarity measures and paraphrase detection techniques, to compare lexical definitions of concepts (Section 3.7). The two components are combined into a hybrid measure, obtaining high cognitive plausibility.

3. Geo Relatedness and Similarity Dataset (GeReSiD) (Section 6.2): a human-generated gold standard for semantic similarity and relatedness. This dataset allows the assessment of the cognitive plausibility of measures when applied on OSM concepts, determining empirically to what degree a measure approximates relatedness and similarity. GeReSiD overcomes the limitations of existing similarity datasets.

4. The *similarity jury* (Section 5.7): the combination of similarity measures can obtain better cognitive plausibility than individual measures. This novel approach is analogous to the combination of disagreeing expert opinions, and provides a widely-applicable technique to compute plausible measures of geo-semantic similarity.

5. A holistic, viewport-based Geographic Information Retrieval (GIR) (Section 3.9): this system treats map viewports as documents, and compares them based on their semantic content. Thus, the user can retrieve map viewports that are semantically similar to a given viewport. This approach to similarity offers an alternative mode of exploration of geographic information.

These aspects of our contribution are described in detail in this thesis. Part of this body of work has been published in peer-reviewed international journals and conferences [19, 20, 21, 22, 25, 23, 26, 24, 200]. The next section outlines the organisation of this thesis.

1.5 Thesis structure

The structure of this thesis follows an organisation commonly found in the scientific literature, presenting the material in an introduction, related work, approach, implementation, evaluation, and conclusions. The current chapter has introduced the motivation, aims and contribution of this thesis. Chapter 2 reviews in detail the related work, starting from user profiling and personalisation (Section 2.2). Several research areas inform our contributions to OSM semantics, starting from the emerging area of spatial crowdsourcing (Section

2.3). The area of geo-semantics draws from a broad scientific literature, ranging from linguistics, to computer and cognitive science (Section 2.4). Interdisciplinary research on semantic similarity is then surveyed, identifying issues relevant to the context of OSM geographic concepts (Section 2.5).

Chapter 3 outlines the approaches we have developed as a contribution to providing semantic support tools for OSM. The RecoMap system highlights geographic features based on users' behaviour. In order to semantically enrich the OSM dataset, we devise a technique to enrich the OSM dataset with DBpedia and LinkedGeoData (Section 3.2). Subsequently, we present a novel open-source resource, the OSM Semantic Network (Section 3.3), aiming at providing support for 'stronger' semantics (Section 3.4). The OSM Semantic Network can be used to compute the semantic similarity and relatedness of geographic concepts (Section 3.5). Our approach to semantic similarity for OSM concepts, NetLexSiM, is based on two complementary components: network-based, co-citation topological measures (Section 3.6), and a knowledge-based measure to compare the lexical definitions of concepts (Section 3.7). These are combined in a hybrid similarity measure (Section 3.8). A holistic, viewport-based GIR system explores an alternative approach, treating map viewports as holistic semantic units (Section 3.9).

Technical aspects of the implementation of these approaches are detailed in Chapter 4. The OSM Semantic Network is extracted by an open source semantic tool that we have implemented, the OSM Wiki Crawler (Section 4.2). The chapter then focuses on our Web platform for map personalisation and visualisation, developed as part of our preliminary exploration into map personalisation and implicit feedback. While investigating suitable technologies for the development of such a platform, we conducted a survey, asking open source contributors their opinion about a set of Web mapping tools (Section 4.3). The resulting Web platform for map personalisation and visualisation hosts Web services tailored to exploit implicit feedback mechanisms to provide a range of personalisation services (Section 4.4).

Chapter 5 reports an empirical evaluation of the aforementioned approaches, starting from our technique to enrich the semantics of OSM (Section 5.2). To evaluate the effectiveness of similarity measures for OSM, we take their cognitive plausibility into consideration (Section 5.3). To conduct a pilot evaluation, an existing gold standard is utilised (Section 5.4). On this dataset, a pilot evaluation of network-based algorithms is performed, reaching high cognitive plausibility for the algorithm SimRank (Section 5.5), followed by the lexical similarity (Section 5.6), including the analogy of the *similarity jury* (Section 5.7). A full evaluation of these techniques, tailored for the OSM semantic model, is described in Chapter 6. To further evaluate our approach, we developed the Geo Relatedness and Similarity Dataset (GeReSiD) (Section 6.2), and used it to evaluate network (Section 6.3), lexical (Section 6.4) and hybrid similarity (Section 6.5), obtaining the highest cognitive plausibility with the hybrid approach. A preliminary evaluation of our approach to holistic viewport similarity concludes the chapter (Section 6.6).

Chapter 7 draw conclusions on the research efforts presented in this the-

sis. After summarising the structure of the thesis (Section 7.1), we discuss its original scientific contribution (Section 7.2). The limitations of the various approaches are discussed, indicating possible solutions (Section 7.3). Finally, we point out directions to conduct future research in the area of map personalisation, VGI, and semantic similarity and relatedness (Section 7.4), showing how this work will be further extended in the near future.

CHAPTER

TWO

RELATED WORK

2.1 Overview

The body of work described in this thesis involves aspects of spatial crowdsourcing, Volunteered Geographic Information (VGI), geo-semantics, and semantic similarity. This chapter has the purpose of surveying the state-of-the-art in these complex, active research areas, framing our efforts in a broad scientific context. The lack of a formal semantic structure in OpenStreetMap (OSM) is a barrier to the effective usage of its vast, rapidly evolving vector dataset. In order to frame our contribution to filling the semantic gap between VGI and the Semantic Geospatial Web, this chapter surveys these areas of research. Our contribution, as Isaac Newton and many before him pointed out, stands on the shoulders of giants.

The work presented in this thesis has been developed to provide semantic support for map personalisation, user profiling, and automatic collection of implicit feedback (Section 2.2). The explosive change in online geographic information is not only quantitative, but also qualitative. Spatial crowdsourcing has triggered a nexus of interrelated phenomena, bringing VGI to the forefront (Section 2.3). A major challenge in this new digital landscape is the interpretation, integration, and effective usage of such diverse and noisy geo-data sources. Semantics represents the cornerstone for generating intelligible and usable data (Section 2.4). Several approaches to semantics exist, originating from logic, linguistics, and Geographic Information Science (GIScience).

The human ability to assess similarity plays a central role in cognition and, therefore, in semantics. To solve a wide range of cognitive and semantic problems, techniques to quantify similarity have been proposed (Section 2.5). An important aspect of our contribution is the exploration of semantic similarity techniques for crowdsourced geographic concepts in the OSM Semantic Network, which we have extracted from the OSM Wiki website. To compute concept similarity in this crowdsourced semantic network, co-citation network measures are promising. Furthermore, lexical semantic similarity offers another semantic dimension to analyse. These measures are not just right or wrong, but can be considered cognitively plausible to the degree to which they approximate human judgement. For this reason, several similarity gold standards have been created and utilised as a cognitive ground truth. Finally, as part of our work on semantic similarity, we investigate the work related to

the computation of the semantic similarity of viewports, and not of individual geographic concepts or features.

2.2 User profiling and personalisation

The work on OSM semantics in this thesis aims at filling a conceptual gap in its semantic model, and was conceived while conducting work on map personalisation and implicit feedback. In the context of Web spatial applications, an approach to improving user experience consists of user profiling and personalisation. These research areas have attracted huge interest both in academia and industry. Understanding users' behaviour constitutes the core of several interface personalisation techniques, aimed at increasing the relevance and accessibility of information for human users. Over the past decade numerous approaches to filter out irrelevant information from large data sets through user profiling have been proposed [202, 203, 6]. One of the typical applications based on user profiling is that of recommender systems, which aim at proposing relevant items to the users from corpora too large to be scanned manually.

Balabanovic and Shoham [18] have defined as (1) *content-based* a recommendation based solely on a profile built up by analyzing the content of items which a specific user has rated in the past, whilst a (2) *collaborative* recommendation relies on a similarity measure among user profiles, without taking the item content into account. Several recommender systems adopt a (3) *hybrid* approach, combining content-based and collaborative recommendations to overcome the limitations of each approach [235, 5].

Collaborative profiling has major benefits [259]. For example, by defining a user profile similarity function it is possible to cluster users who have similar interests and add new elements to the interest score calculation. However, despite the extensive body of research in web personalisation and recommendation, comparatively little work has been carried out on the spatial and geographic dimension of data, whose commercial potential has become apparent in recent years [267]. To provide content personalisation and recommendation, the knowledge about the user needed for personalisation can be inferred from *interest indicators*, meaning pieces of information about the user preferences and behaviour. Interest indicators are behaviours such as clicking on a link, copying and pasting text in a document, indicating in a form an interest in photography, sharing an article with a colleague, visiting a Web site regularly, and so on. These feedback indicators can be collected in two ways, explicitly and implicitly.

In the case of (1) *explicit* feedback, the system asks the user to enter preferences and to rate objects manually (e.g. the *like* buttons on the Amazon website). The feedback collected this way is generally accurate but is time consuming and distracts the user from their main task [168, 312]. Moreover, some users do not change programme preferences and skip feedback forms, so it might be difficult to collect information about them. On the other hand, (2) *implicit* feedback aims at gathering knowledge about a user (or a group of users)

by considering as indicators aspects of their interaction with the system. The main problem with this approach is that the extraction of valuable information from a large amount of recorded interaction can be difficult, as users can interact with things by mistake, or can suddenly change task, and so on. Kelly and Teevan have surveyed this area of research, providing an account of the various implicit feedback indicators that have been utilised and their degree of reliability [147].

Several Web-based systems employ an implicit profiling approach based on indicators such as link clicking, page printing, e-mailing, etc. This approach has also been developed in the geo-spatial domain based on user interactions with map contents [305, 187, 37]. The system CoMPASS is a GIS application that monitors user interaction to recommend groups of features, such as layers, to users [305]. Similarly, the system described by Mac Aoidh and Bertolotto [187] registers user mouse interactions to elicit interests, with the intention of providing recommendations at the object level. Hiramoto and Sumiya [123] have defined 'operation chunks' to interpret the user behaviour on an interactive map. Typical operations are 'narrowing-down', 'spreading-out', 'panning', and 'relative-position confirming'. Such spatial implicit feedback has been subsequently used to personalise Web maps, and location-based services (LBS) [154, 282]. Moreover, Kotera et al. [159] devised a system that recommends regions and geographic classes to users, based on their interaction history.

To provide effective personalisation, data semantics plays a crucial role. Finding meaning in unstructured - or loosely structured - online information has been the main challenge addressed by the set of initiatives under the name *Semantic Web* at the beginning of the millennium [31]. After a decade, ontologies are still considered one of the key elements of the Semantic Web and are technologically supported by languages such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL). Ding et al. [61] provide a detailed survey of the field of ontologies for the Semantic Web.

Personalisation is one of the fields in which ontologies have been utilised to build user profiles for location-based services [316, 111], to develop context-aware mobile systems [158] and to refine Web searches through ontological user profiles [268]. Although ontologies have a long and well-established scientific tradition, little work has been done in the area of spatial implicit feedback analysis. In order to gain a better understanding of user spatial interests, this linkage between raw vector data and knowledge structures, such as ontologies and semantic networks, can be beneficial. For example, a user that interacts very often with objects of type 'school' and 'university' is likely to have an interest in 'education', and her user profile should model this piece of information.

Implicit feedback indicators, such as mouse movements and map navigational behaviour, can be used to infer user interests [188]. The explicit semantic associations provided by LinkedGeoData between the map presented to the user and its underlying ontological entities represent a step in this direction. Once these ontological relationships are established, it is possible to combine

them with implicit indicators to enrich user profiles, capturing the semantics of their interests in a more sophisticated way. But, in order to infer valid user interests from implicit feedback indicators such as mouse movements, a more detailed and comprehensive mapping between spatial data and ontological data is needed.

Our contribution to the area of spatial recommender systems consists of RecoMap (see Section 3.2.1). RecoMap monitors the user location to emphasise geographic features on the map. The RecoMap prototype was implemented as a part of a Web platform for map personalisation and visualisation (see Section 4.4). Our contributions to geospatial semantics provide semantic support for implicit feedback analysis and map personalisation. The next section moves to the core of this thesis, the rise of spatial crowdsourcing and VGI.

2.3 Spatial crowdsourcing

The last decade has witnessed rapid changes in the production and consumption of geographic information. In this period, the traditional GIS field has progressively become Web-based and universally accessible through Web technologies [114]. Once collaborative tools reached a certain maturity, online efforts have started, resulting in several crowdsourced spatial projects, adding a new element towards the utopian vision of a full digital coverage of the Earth [97]. This transformation of users from passive end points of services to active contributors has wide ranging consequences for Geographic Information Systems (GIS) and, ultimately, also for GIScience.

2.3.1 Volunteered Geographic Information (VGI)

Digital geographic information has experienced an unprecedented growth during the past decade, both in quantitative and qualitative terms. *Neogeography* is the umbrella term that Turner coined in his 2006 study on online mapping tools and resources, to refer to this nexus of phenomena [291]. Michael Goodchild [96] termed the crowdsourcing of geographic information as Volunteered Geographic Information (VGI), specifically referring to geographic information produced and released by non-expert users through *voluntary* actions. Goodchild [97] has criticised the term 'neogeography', noting that geography is an established scientific discipline, while VGI phenomena rely on amateur contributors. VGI is having a visible impact on the production and consumption of geographic information, adding a collaborative dimension to the traditionally hierarchical, centralised model of production [114, 51, 67].

In parallel, the expansion of mapping to increasingly powerful mobile computing devices has led to the so-called 'ubiquitous cartography' [89]. Sui [281], discussing neogeographic trends, suggests the term 'wikification' to capture the attempt to crowdsource non-textual data, emulating Wikipedia within the spatial domain. In this context, Priedhorsky and Terveen [237] proposed an adaptation of the wiki model for spatial data. The growth of available online

geographic information raises the issue of semantics: data is useless unless its meaning is intelligible. The threat of a deluge of semantically poor geo-data prompted Egenhofer [65] to envisage the emergence of the Semantic Geospatial Web, a spatial extension of the Semantic Web initiative that will enable advanced information retrieval.

The impact of VGI is not restricted to non-profit, academic organisations. Private data providers are progressively offering facilities for sharing geo-data, expanding their services beyond the routing systems that dominated the last decade. Geo-wikification is identifiable in the growth of web services allowing users, with some degree of freedom, to create or edit spatial data. As Priedhorsky and Terveen [237] noted, however, most interactive geo-services are essentially 'digital graffiti,' i.e. annotations on a static geographic image.

Overall, neogeography can be defined as crowdsourced, wikified, interactive, web-based, volunteered, and ubiquitous. Several neogeographic online projects gathered wide communities of users and contributors. These projects range from the very specific – Cyclopath collects cycling-related geographic knowledge – to the very generic – Wikimapia is a commercial effort to build an online editable map where the user 'can describe any place on Earth'.[1] The next section surveys the existing online projects that generate open geo-knowledge bases.

2.3.2 Open geo-knowledge bases

This section provides a survey of open, collaborative geo-knowledge bases, which constitute an important part of semantic technologies. As a result of the synergy between crowdsourcing, VGI, and the Semantic Geospatial Web, several large-scale collaborative projects have emerged. Several geo-knowledge bases have then been created by structuring existing datasets into Semantic Web formats: the projects LinkedGeoData, GeoNames, and GeoWordNet are particularly significant examples [15, 92]. Research efforts have been undertaken on the development, maintenance, and integration of open geo-knowledge bases, to enhance the geographic intelligence of information retrieval systems, beyond the traditional text-based techniques. This survey was published in [26].

To avoid terminological confusion, it is beneficial to provide a definition of the related and sometimes overlapping terms used in knowledge representation. A 'knowledge base' is a collection of facts about a domain of interest, typically organised to perform automatic inferences [109]. A knowledge base contains a terminological conceptualisation (typically called 'ontology') and a set of individuals. Widely used both in philosophy and in computer science, the meaning of the term 'ontology' is particularly difficult to define [272]. Among the many definitions, 'an explicit specification of a conceptualization' and 'shared understanding of some domain of interest' are of particular relevance, as they stress the presence of an explicit formalisation, and the general

[1] `http://cyclopath.org`, `http://wikimapia.org`

aim of being understood within a given domain [109, p. 587]. Winter notes that ontologies became part of GIScience towards the end of the 20th century [308].

A 'thesaurus' is a list of words grouped together according to similarity of their meaning [251], whilst a digital 'gazetteer' is specifically geographic, and contains toponyms, categories, and spatial footprints of geographic features [122]. In the Web 2.0 jargon, a 'folksonomy' is a crowdsourced classification of online objects, based on an open tagging process [297]. Finally, a 'semantic network,' a term which originated in psychology, is a graph whose vertices represent concepts, and whose edges represent semantic relations between concepts [239].[2] We define a 'geo-knowledge base' as a knowledge base containing some geographic information.

The Semantic Web and the Linked Open Data initiative promote the adoption of semantic formats, which can be used to add an open, machine readable semantic structure to online data [33]. In this context, several collaborative projects have emerged, resulting in a growing number of freely available geo-knowledge bases. This section provides a survey of such open geo-knowledge bases, restricting the scope to projects having global coverage, discussing their spatial content. These artifacts are the result of combined efforts in crowdsourcing, VGI, and the Semantic Geospatial Web, and offer useful resources for Geographic Information Retrieval (GIR), and other areas of geoscience. Most projects discussed in this section appear as hybrid, dynamic collections of heterogenous knowledge, including aspects of gazetteers, folksonomies, and semantic networks.

Among these numerous resources, we focus on eleven datasets that have a global scope (as opposed to local projects), are mostly generated through crowdsourcing, released under Creative Commons/Open Database licences,[3] and are available as fully downloadable dumps in popular Semantic Web formats such as OWL and RDF. Some of the selected projects are focused specifically on geographic data (e.g. GeoNames and OpenStreetMap), while others are more general-purpose but contain valuable geographic knowledge (e.g. DBpedia and Freebase). These knowledge bases provide open datasets, and are strongly inter-connected with one another. Our own contribution to this area of research, the OSM Semantic Network, is described in Section 3.3. Relevant characteristics of each project are summarised in Table 2.1.

CONCEPTNET This semantic network is focused on natural language processing and understanding [120]. ConceptNet is a large semantic network, whose nodes represent concepts in the form of words or short phrases of natural language. The graph edges represent labeled relationships. Each statement in ConceptNet has justifications pointing to it, explaining where it comes from and how reliable the information seems to be. The network includes 1.6 million assertions gathered from Wikipedia,

[2] Unlike ontologies, semantic networks focus on psycho-linguistic aspects of the terms. However, some artifacts, such as WordNet, defy this distinction by showing aspects of ontologies and semantic networks.

[3] `http://creativecommons.org`, `http://opendatacommons.org`

Project Name	*Year**	*Type & Content*	*Data sources*	*Formats*
CONCEPTNET	2000	Semantic network, knowledge base; 1.6 million assertions, 700,000 natural language sentences	Wikipedia, WordNet, and others	JSON
DBPEDIA	2007	Knowledge base; 320 classes, 740K Wikipedia types, 3.6M entities, 1 billion triples	Wikipedia	OWL/RDF
FREEBASE	2007	Knowledge base; 22M+ entities, 1M locations	Crowdsourced	Tab separated text
GEONAMES	2006	Gazetteer; 650 classes, 10M+ toponyms	Gazetteers, Wikipedia, crowdsourced	OWL/RDF
GEOWORDNET	2010	Semantic network, thesaurus, gazetteer; 330 classes, 3.6M entities	WordNet, GeoNames, MultiWordNet	RDF
LINKEDGEODATA	2009	Gazetteer; 1K classes, 380M geographic entities	OpenStreetMap	RDF
OPENCYC	1984	Semantic network, knowledge base; 50K classes, 300K facts	Expert-authored Cyc knowledge base	OWL/RDF
OPENSTREETMAP	2004	Vector map, gazetteer; User-defined tags, 1.2B nodes, 114M ways	Crowdsourced, free GIS datasets	XML
WIKIPEDIA	2001	Semantic network, dictionary, thesaurus; Semi-structured (infoboxes), 3.9M articles in English	Crowdsourced	XML
WORDNET	1985	Semantic network, dictionary, thesaurus; 117K synsets	Expert-authored knowledge base	OWL/RDF
YAGO	2006	Knowledge base, semantic network; 10M+ entities, 460M facts	Wikipedia, GeoNames, WordNet	RDF

Table 2.1: A survey of open geo-knowledge bases. All of these projects are currently active, release open data, have global scope, and are interconnected with other projects. *Beginning of the project.

Wiktionary, WordNet, and 700,000 sentences from the Open Mind Common Sense project [270]. Efforts to encode ConceptNet in RDF are being undertaken [103].

DBPEDIA One of the leading projects of the Semantic Web, DBpedia is a Semantic Web version of Wikipedia [14]. The knowledge base currently contains 3.6 million entities, including 526,000 places, encoded in a billion RDF triples. As DBpedia is strongly interconnected with other geo-knowledge bases (e.g. WordNet W3C, GeoNames, LinkedGeoData), it is considered the central hub of the Linked Open Data.

FREEBASE Designed as an open repository of structured data, Freebase allows web communities to build data-driven applications [35]. The dataset is structured around terms (classes), and unique entities (instances), where an entity can be a specific person, a place, or a thing, and is described by facts. It currently contains 22 million entities, of which 1 million are locations. As entities are described by facts corresponding to a directed graph, it can be easily converted into RDF.

GEONAMES Combining multiple data sources, GeoNames aims at offering a large, volunteered gazetteer.[4] The project contains over 10 million toponyms, structured in 650 classes. GeoNames integrates geographical data such as names of places in various languages, elevation, and population. The data is collected from traditional gazetteers such as those of the National Geospatial-Intelligence Agency (NGA) and the U.S. Geological Survey Geographic Names Information System (GNIS), and is crowdsourced online.

GEOWORDNET GeoWordNet is the result of the integration of WordNet, GeoNames and the Italian part of MultiWordNet [92]. It is a hybrid project, combining a semantic network, a dictionary, a thesaurus, and a gazetteer. It was developed in response to the limited WordNet coverage of geospatial information and the lack of concept grounding with spatial coordinates. The semantic network contains 3.6 million entities, 9.1 million relations between entities, 334 geographic concepts, and 13,000 (English and Italian) alternative entity names, for a total of 53 million RDF triples.

LINKEDGEODATA (LGD) Since OSM consists of a large collection of geographic data, LinkedGeoData is an effort to republish it in the Semantic Web context [15]. The OSM vector dataset is expressed in RDF according to the Linked Open Data principles, resulting in a large spatial knowledge base. The knowledge base currently contains 350 million nodes, 30 million ways (polygons and polylines in the OSM terminology), resulting in 2 billion RDF triples. Some entities are linked with the corresponding ones in DBpedia.

[4] `http://www.geonames.org`

OPENCYC This is the open source version of Cyc, a long running artificial intelligence project, aimed at providing a general knowledge base and common sense reasoning engine.[5] Even though OpenCyc covers a limited number of geographic instances, it contains a rich representation of specialised geographic classes, such as *salt lake* and *monsoon forest*. The OpenCyc classes are inter-linked with DBpedia nodes and Wikipedia articles.

OPENSTREETMAP (OSM) The OSM project aims at constructing a world vector map [113]. The leading VGI initiative, the dataset represents the entire planet, gathering data from existing datasets, GPS traces, and crowdsourced knowledge. To date, the vector dataset contains 1.2 billion nodes (points), and 115 million ways (polygons and polylines). The project is described in detail in Section 2.3.3.

WIKIPEDIA A collaborative writing project, Wikipedia is a multilingual, universal encyclopedia, and has become the most visible crowdsourcing phenomena. The English version currently contains 3.9 million articles, resulting in a 2 billion word corpus. Because of high connectivity between its articles, Wikipedia is sometimes used a semantic network [279]. This vast repository of general knowledge has been used for different purposes, including semantic similarity and ontology extraction [299, 218]. The project has attracted interest in the area of GIScience [7].

WORDNET Initially conceived as a lexical database for machine translation, WordNet has become a widely used resource in various branches of computer science, where it is used as a semantic network [74]. Currently it contains 117,000 'synsets', groups of synonyms corresponding to a concept, connected to other concepts through several semantic relations. The dataset has been encoded and released in RDF, becoming a highly linked resource in the web of Linked Open Data.[6] Details of the WordNet semantics are discussed in Section 2.5.4.

YAGO Yet Another Great Ontology (YAGO) is a large knowledge base extracted from Wikipedia and WordNet [280]. Recently YAGO has been extended with data from GeoNames, with particular emphasis on the spatial and temporal dimensions [125]. The current version of the knowledge base contains 10 million entities, encoded in 460 million facts. YAGO is inter-linked with DBpedia and Freebase.

Figure 2.1 depicts this constellation of open geo-knowledge bases, showing a schematised data path from the data producers to the knowledge bases. Bearing in mind the complexity of these collaborative processes, the main actors in this constellation, involved in the production of information and the generation of open linked data, can be grouped as follows:

[5] `http://www.cyc.com/opencyc`, `http://sw.opencyc.org`
[6] `http://www.w3.org/TR/wordnet-rdf`

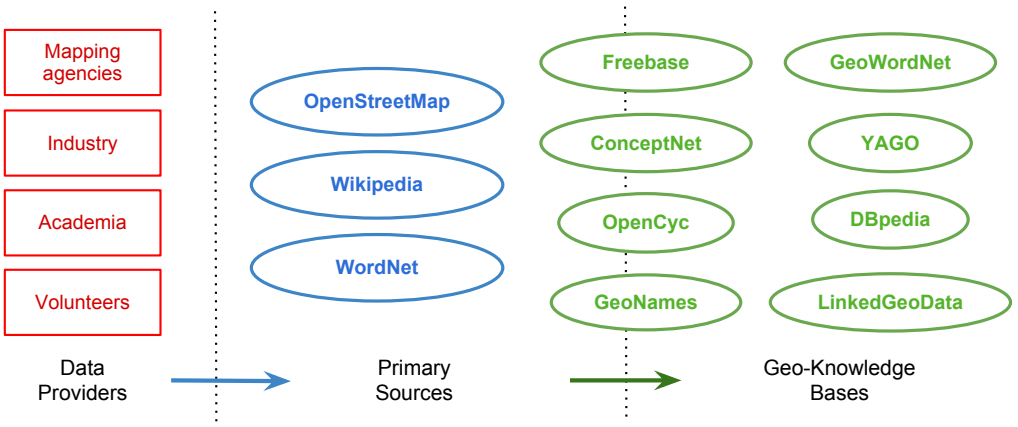

Figure 2.1: The constellation of open geo-knowledge bases. The data path is schematised from the data providers to semi-structured primary sources, and finally structured into knowledge bases. Some projects defy classification by producing new knowledge in structured knowledge bases, and extracting knowledge from primary sources.

1. **Data providers.** Traditionally, geographic data was collected exclusively by experts and professionals in large public and private institutions. As Web 2.0 and VGI have emerged, a new category of non-expert users/producer ('produsers') has entered the production process [51]. Crowdsourced primary sources include contributions from a wide variety of information producers, ranging from experts operating within public and private institutions to non-expert, unpaid, pro-active users.

2. **Primary sources.** Projects such as Wikipedia and OSM collect a large amount of information about the world through crowdsourced efforts. On the other hand, primary sources such as WordNet are expert-authored, while other projects combine both crowdsourcing and expert control. Most geo-knowledge bases rely heavily on these primary sources, often aligning and merging them into larger knowledge bases. Inconsistencies and contradictions in primary sources can be propagated to the derived knowledge bases. For example, an incorrect piece of information in a Wikipedia article will be also found in DBpedia and YAGO.

3. **Geo-knowledge bases.** Typically, open geo-knowledge bases consist of structured and aggregated versions of existing semi-structured or unstructured primary sources. However, some datasets lie at the boundary between primary sources and knowledge bases, as they are both interlinked with existing knowledge bases and produce new data through crowdsourcing and expert contributions (e.g. Freebase and OpenCyc). Several knowledge bases encode the same primary data into different formalisms, such as DBpedia and YAGO.

These three actors are part of an open system, in which more or less structured data flows in complex patterns that determine the nature, quality and limitations of the resulting projects. Investigations of such collaborative open

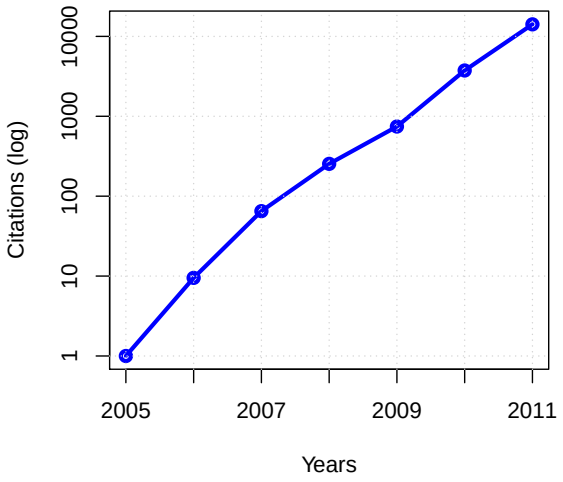

Figure 2.2: Cumulative number of citations to OpenStreetMap from 2005 to 2011. The y axis is logarithmic. Data collected from Google Scholar and Scirus.

processes have been carried out, both in the area of general crowdsourcing and VGI [51]. In order to contribute to OSM semantics, we have developed a new resource, the OSM Semantic Network (see Section 3.3). The next section focuses on the leading VGI project.

2.3.3 OpenStreetMap (OSM)

Leading grassroots mapping project, OpenStreetMap (OSM) aims at creating an open vector map of the world [113]. Unlike other VGI projects, OSM revolves around the construction of a vector dataset representing the entire planet, not just annotations on an existing map, and emphasises the openness of its datasets.[7]

Since British computer scientist Steve Coast founded it in 2004 [49], the project has increasingly attracted attention both in industry and in academia. As a rough but significant indicator of the interest aroused by OSM in academia, we examined the number of scientific publications that reference the project. Figure 2.2 shows that the interest in OSM has grown exponentially over the past 6 years. This citation count was obtained by averaging the counts collected from the scientific search engines Google Scholar via the 'Publish or Perish' tool [119] and Scirus.[8] The project is experiencing an increasing success in industry too: among many other companies, the spatial social network Foursquare recently started a transition from Google Maps towards the adoption of OSM data [115].

Given the project's reliance on the Wiki model, one of the most dis-

[7] http://wiki.openstreetmap.org/wiki/OpenStreetMap_License
[8] http://scholar.google.com, http://www.scirus.com

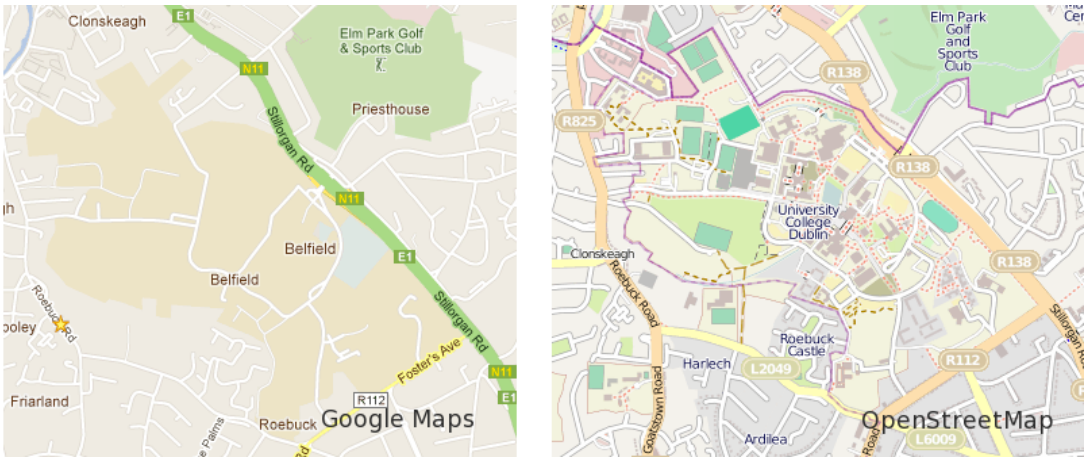

Figure 2.3: Google Maps (left): Map data ©2012 Google, Tele Atlas; OpenStreetMap (right): Map data CCBYSA 2012 OpenStreetMap.org. Images extracted on March 17, 2012.

cussed issues is data quality, which for the moment remains an open problem [112, 212]. Haklay [112] has conducted a preliminary investigation comparing OSM with British Ordnance Survey data, noting that "OSM information can be fairly accurate" (p. 682). Similarly, Mooney et al. [212] are working towards a quality metric for OSM. Although agreement on a formal assessment of the data quality has still to be reached, OSM can at times outperform commercial and traditional data providers in terms of informational richness. Figure 2.3 shows an example of this phenomenon, comparing OSM and Google Maps on the area of Dublin where the University College Dublin campus is located.

While the geometric quality of OSM data is debated, little work has been done on the *semantic* quality of the classes under which the geometries are classified. The OSM geometries are described through tags, which indicate their meaning and functional role in the dataset. For example, a university campus consists of a polygon delimiting its boundaries, associated with the tag *amenity=university* (see Section 3.3 for a detailed discussion of OSM semantic model). A set of similar and sometimes competing projects has emerged to enhance OSM semantics. LinkedGeoData has taken the entire OSM dataset and republished it in a Semantic Web-friendly format, linking it to a taxonomical ontology that contains 1,293 classes [15]. Despite the advantages of the new format, the LGD ontology is a simple, shallow taxonomy representing tags as key and values. Its semantic content is limited to is_a relationships between tags and respective values. OSMonto[9] offers another ontology based on OSM tags. Its main dataset consists of an incomplete formal description of a subset of OSM tags.

To date, none of these projects has been officially integrated into the project infrastructure, and the OSM semantics has been largely left unexplored. Furthermore, to the best of our knowledge, none of the aforementioned projects provides a semantic similarity measure for OSM geographic concepts. Our core contribution to OSM consists of an expansion of its seman-

[9] `http://wiki.openstreetmap.org/wiki/OSMonto`

tic model, enabling, among other applications, the computation of the semantic similarity of geographic concepts. The next section surveys the core tenets of semantics, a broad interdisciplinary research area.

2.4 Semantics

This section surveys the main semantic theories that have dominated the debate in linguistics, cognitive science, and GIScience. Since the inception of the Semantic Web, the word 'semantic' has become ubiquitous in basic and applied research in computer science [31, 65]. A recent development that has given further momentum to semantic technologies is the announcement of the Google 'semantic search' product, aimed at extending keyword-based information retrieval (IR) with ontologies, and advanced knowledge representation [64]. This interdisciplinary survey frames our work in the broader context of the formal study of meaning, crucial to understand the great challenges that semantic geo-technologies have to face.

It is uncontroversial that humans have the remarkable ability to convey meaning through language, creating, sharing and manipulating abstract concepts, beyond the constant flow of irrelevant details provided by the senses. The inherent difficulties of theorising about meaning is the fact that language is a tool to describe extra-linguistic phenomena, and that meaning 'happens' somewhere between the linguistic and the extra-linguistic reality. In his *Critique of pure reason*, Immanuel Kant [142] noted that, based on the concept of a dog, "my imagination can delineate the figure of a four-footed animal in general, without being limited to any particular individual form which experience presents to me" (p. 83). On the other hand, the formal analysis of the workings of meaning has proven to be, to say the least, very challenging, and can account for the difficulties encountered by the Semantic Web [82].

The study of meaning is one of the time-honoured enterprises of Western philosophy. Plato's discussion on language in his *Cratylus* represents an early attempt to disentangle the mystery of word meaning, pointing out its conventional nature [231]. Although Plato seemed quite skeptical about the possibilities of a complete semantic theory, several arguments and theories have been discussed by subsequent scholars.[10] Broadly speaking, semantics is the area of inquiry that aims at explaining how meaning is created and transmitted in human communication, through informational 'markers' such as words, sentences, scents, sounds, drawings, gestures, etc. In this sense, while information theory focuses on the *transmission* of information, semantics focuses on how the information is transformed into *meaning* for the human sender and receiver [265]. The main concerns of semantics are located in the relationships between these markers and abstract concepts, between humans and markers, and between markers. The formal study of semantics is a core preoccupation in philosophy, logic, and linguistics [90].

[10] Among others, notable contributors are Aristotle [209], Frege [84], Tarski [285], Wittgenstein [309] and, more recently, Katz [144].

The following sections briefly survey semantic theories in logic (Section 2.4.1) and linguistics (Section 2.4.2). Given the spatial aspect of the OSM concepts, Section 2.4.3 explores semantics in the context of geographic information science.

2.4.1 Logical semantics

This branch of semantics developed as part of logic and analytic philosophy of language in the 19th and 20th century, an area dominated by German logician Glottlob Frege [243]. Focusing on the relationship between language and reality, such semantic theories purport to explain the meaning of propositions in terms of necessary and sufficient conditions, to obtain truth-conditions through logical rules.

The so-called 'theory of reference' is a milestone of philosophical semantics [275]. The basic tenet of this approach is that the meaning of propositions is to be found in their 'reference' to some existing, stable entity in the world. For example, the meaning of the term 'River Liffey' consists of its reference to an actual river located in Ireland. In this regard, Frege put forward an influential distinction between 'sense' (*Sinn* in German), and 'reference' (*Bedeutung*). In a proposition, Frege argued, the 'reference' is what the proposition refers to, whilst the 'sense' is how the proposition describes its reference. This way, two propositions can have the same reference and yet different senses. Sense, in other words, represents a specific viewpoint from which the reference is being observed, and determines the reference.

A related and important distinction, first developed by Gottfried Wilhelm Leibniz, is that between intensional and extensional meaning of a word [278]. The 'intensional meaning' is the set of attributes describing the object. On the other hand, the 'extensional meaning' of a sign is generally considered as the set of objects in a possible world that are referred to by the sign. For example, the intensional meaning of the noun 'volcano' can be 'a conical hill or mountain, having a crater or vent', 'a crack in the Earth's crust from which hot gases and lava are released', etc. Its extensional meaning would embrace all actual objects matching the intensional meaning, such as Krakatoa and Vesuvius.

In this context, it is worth pointing out the distinction between 'semantics' and 'pragmatics.' The term semantics tends to be restricted to *constant* properties of terms, while pragmatics refers to aspects that vary from context to context. For example, the term 'local shop' has the same meaning regardless of when I use it, but its extension varies from one context to another: the extension of 'local shop' is pragmatically determined. On the other hand, the extension of 'lake' does not vary from one context of use to another [197]. However, there is little agreement on the precise theoretical scope of these terms, and the debate lies beyond the scope of this thesis [43].

Later qualifications notwithstanding, referential theory fails to account for the enormous role played by subjectivity in language, adopting a view of meaning as objectively existing between propositions and the material world.

In the 20th century, semantic theories developed by linguists overcame these drawbacks, proposing a substantial change of viewpoint. In computer science, logic-based semantics is crucial to fields such as artificial intelligence (AI) and the Semantic Web [36]. The fundamental limitations of this approach to semantics helps grasp the difficulties that such technological enterprises have been encountering.

2.4.2 Linguistic semantics

The meaning of linguistic markers has been thoroughly explored in modern linguistics and semiotics. In this context, Geeraerts [90] gives an up-to-date and exhaustive account on the major theories of lexical semantics, while Maienborn et al. [191] offer a wide-ranging survey of modern semantics and lexical meaning.

Linguistic semantics originated from philology, attempting to explain how the meaning of individual words can change over time, and what is likely to cause these shifts.[11] In the early 20th century, Swiss linguist Ferdinand de Saussure proposed the so-called structuralist semantics, marking a change in the scope and methods of semantics [90]. Rather than as a set of independent words, whose meaning can be determined in isolation, language was to be seen as a system of 'signs,' i.e. inter-related arbitrary symbols. In structuralist semantics, signs cannot be observed in isolation, but always in relation to other signs. Saussure famously drew an analogy between language and chess, as arbitrary symbolic systems in which the value of a unit makes sense only when considering it as part of the system [258].

In Saussure's theory, a 'sign' is made of a signifier and a signified. A 'signifier' is a conventional symbol that refers to a target object. Utterances, written symbols, body gestures, can be considered to be signifiers. The target of a signifier is a 'signified,' i.e. the thought of a physical or an abstract concept, such as a tree, a lake, or the idea of movement in space. Communication between human beings happens through the sharing of signifiers. Language, in structuralism, can be seen as an autonomous, intermediate layer between the human mind and the outside world, a network of signifiers. Understanding semantics, therefore, should hinge on the study of the structural architecture of this layer. While a signified can indeed have a referent in the real world, Saussure considered the issue to be completely outside of the scope of semantics. Ogden and Richards [225] criticised this position, and proposed a structuralist semantic theory that stressed the important of actual things in the world, the referents. Their renowned semiotic triangle outlines the relations between the 'thought or reference' (Saussure's signified), the 'symbol' (signifier), and the 'referent'.

[11] The term, from ancient Greek *sema*, 'sign', was coined by French linguist Michel Bréal [90]. Key figures were Max Hecht, who defined the essential nature of historical-philological semantics, and Michel Bréal, who stressed the role of the semantician as being "to interpret historical texts against the background of their original context by trying to recover the original communicative intention of the author" [90, p. 14].

An influential structuralist method to deal with word meaning, which informs our definition of semantic relatedness, was proposed by Jost Trier in the 1930s. Trier formulated his 'lexical field theory' as a set of semantically related words describing a restricted domain of human knowledge [90]. To illustrate his approach, Trier compared words to the contiguous stones forming a bi-dimensional mosaic. Initially very successful, this theory suffered from demarcational problems, i.e. lexical fields are fuzzy, and boundaries are difficult to draw not only between different lexical fields, but also within the same field.

Another major difficulty for structuralist semantics is inherent to the core tenet of the approach. Although semantic relationships are essential to the construction of meaning, they cannot fully capture the richness and complexity of word meaning. Concepts can be isolated and defined in the abstract, but when applied to real life situations, the unescapable complexity of reality makes the mapping between signifiers and signified more ambiguous. As Geeraerts [90] pointed out, "the mind is neat but the world is fuzzy" (p. 95). Moreover, the autonomy of language in structuralism turned out to be a strong limitation, and was overcome by cognitive semantics in the second half of the century.

In the 1960s, syntax was at the core of Noam Chomsky's efforts to develop a general theory of language [46, 90]. A branch of semantics reacted to this syntax-dominated research programme, showing a novel interest in the psychological dimension of meaning [144]. This novel approach was based on the assumption that language is not, as structuralists thought, an autonomous system, but is deeply intertwined with human cognition. From the 1980s to date, cognitive semantics is the most popular approach, guiding the efforts in the area of the study of meaning [90].

According to cognitivist semanticians, semantics has to be analysed in the relationship between language and cognitive processes, including concept formation, knowledge acquisition, organisation, etc. [163]. Meaning is not a linguistic phenomenon, but is the result of a cognitive process. A central tenet to this approach is that cognition is *embodied* in the physical/perceptive experience of humans [165]. The physical reality that humans perceive through their body shapes the fundamental architecture of language, and therefore of meaning. Therefore, cognitive semantics endorses a belief in the contextual and dynamic nature of meaning [90]. The idea of embodied cognition is currently a major intellectual trend, affecting not only linguistics, but also cognitive science, robotics, psychology, and philosophy [56].

In computer science, linguistic semantic theories have played a major role in offering models and approaches to tackle the complexity of natural language. A visible expression of structuralism is the so-called semantic networks [310, 239]. Such structured collections of concepts, such as WordNet [74] and ConceptNet [120], are largely utilised to describe knowledge through structural relationships. Our OSM Semantic Network falls within this tradition. Section 2.3.2 provided a survey of open knowledge bases, with particular emphasis on geographic aspects. Moreover, the tenets of linguistic semantic theories are often explored in IR, word sense disambiguation, and natural lan-

guage processing [137, 127]. For example, a computable measure of semantic relatedness aims at the identification of a lexical field (e.g. river, lake, and stream).

2.4.3 Geo-semantics

The area of geo-semantics focuses on the workings of meaning in the framework of GIScience. To date, geo-semantics does not appear to be a coherent, unified scientific project, but is rather a collection of different aspects of semantics in GIR, GIScience, in the broad framework of spatial cognition. This section surveys research on geo-semantics, an area that has been experiencing a rapid growth over the past decade.

One of the main reasons why semantics has gained visibility in GIScience is to be found in the increasing availability of geospatial web services. As GI scientist Werner Kuhn [161] has argued, geographic data used to be produced and consumed in local contexts, in which the data semantics was specified in offline manuals. With the rapid emergence of online repositories of geo-data, such offline manuals have been translated into geo-knowledge bases, a major shift in the discipline, resulting in a wide range of incompatible data formats and conceptualisations [308].

The need for a stronger integration of such rich data sources is identified as a key challenge in GIScience: as Goodchild [98] recently put it, "[t]he dream of a single geoportal remains beyond our grasp, and its achievement will rely on improved understanding of the semantics of search and improved approaches to metadata" (p. 97). The need for a common semantic frame of reference, and plausible similarity measures is considered critical. Along similar principles to those inspiring Berners-Lee's Semantic Web, Egenhofer [65] suggested that higher semantic integration would result in enhanced GIR, helping online users to retrieve relevant geo-data, in what he called the Semantic Geospatial Web.

The possibility of grounding semantics in a spatial model has attracted notable interest in cognitive science. The underlying intuition is that of cognitive semantics, according to which meaning is rooted in the human body, and in its relationship with the environment in which it has evolved. In their theory of *conceptual metaphors*, Lakoff and Johnson [164] pointed out the spatial nature of basic metaphors in language. For example, a spatial metaphor such as HAPPY IS UP, resulting in English expressions such as 'I'm feeling up today', is grounded on the fact that drooping posture is expression of sadness or illness, erect posture of positive feelings.

Extending the intuition of spatial metaphors further, it is possible to argue that the inner structure of concepts is spatial. Gärdenfors [88] formulated the theory of 'conceptual spaces.' Rather than representing knowledge using symbolic models or neural networks, concepts can be seen as objects in a multi-dimensional vector space model. Each dimension in the conceptual space represents any quality of the concepts, such as height, having either discrete or continuous values. Thus, concepts are defined as convex regions in

a conceptual space. In such a model, computing the semantic distance (and therefore similarity) between concepts becomes natural.

A bold attempt to ground geospatial semantics into a conceptual space is that of Kuhn's Semantic Reference Systems [160].[12] While it is possible to transform coordinates from a reference system to any other, notes Kuhn, automatic translation between different geographic semantic resources is extremely challenging. To reach interoperability, Semantic Reference Systems (SRS) would offer transformation rules between the geographic conceptual model of two communities, mapping all the concepts to the corresponding ones. This transformation-based approach would avoid the problematic definition of a shared, universal geographic semantics. Although this approach was extremely successful in standardising spatial systems, it is to date unclear how it would be possible to involve diverse communities into a shared semantic system. Considering the infinite possible conceptualisations of the same geographic information, many challenges lie ahead.

A fundamental problem in GIS is the mapping of real-world features into a mathematical model, allowing the human users to manipulate the model in an intuitive way [70]. As Egenhofer and Mark [66] pointed out in their seminal theory of 'naive geography,' humans often tend to construct inconsistent fuzzy models of geographic knowledge, arguing that GISs should match such intuitive, subconscious models. The notion of naive geography is particularly relevant in the emerging context of VGI. While geospatial semantics of professionally produced data is specified in well-defined, top-down ontologies, the scenario of crowdsourced data is far more complex. Geographic features started being described not by rigid formalisms, but by short unstructured segments of free text (i.e. tags), posing an additional challenge for geo-semanticians. OSM is a prominent case of this casual, loose approach to geo-semantics (see Section 2.3.3). As Bishr and Kuhn [32] have argued, semantics is crucial to the successful production and usage of volunteered geographic data sources.

From the perspective of the emerging area of text-based Geographic Information Retrieval (GIR), semantics is crucial to interpret the user interests, and retrieve relevant resources. Geo-semantics offers theoretical and practical approaches to model fuzzy terms used by humans to search spatial information [198]. Users specify their spatial interests in complex and indirect ways, e.g. landmarks instead of locations, metonymical usage of place names, etc., and GIR systems should aim at handling these cases effectively. To date, several ontology-supported approaches have been devised [186, 185, 199, 193]. In summary, the scope of geo-semantics includes, but is not limited to, the following topics:

- Data alignment, ontology alignment, data integration [162]
- Geo-ontology engineering [308]
- Geographic Information Retrieval (GIR) [238]

[12] `http://seres.uni-muenster.de`

– Named Entity Recognition and Classification (NERC) [217]

– Toponym disambiguation [227]

– Spatial reasoning [85]

– Geo-semantic similarity [263]

It is undeniable that to date geo-semantics is a key area of GIScience and GISs. Understanding geo-semantic aspects of the increasingly large spatial information available is crucial to produce it, integrate it with existing datasets, and utilise it effectively in real-world, data-intensive online applications. Our contribution to this field lays in the exploration of measures of semantic similarity and relatedness for OSM concepts. To fill a knowledge gap in geo-semantics, we developed the Network-Lexical Similarity Measure (NetLexSiM), a semantic similarity measure for OSM.

2.5 Similarity

Humans display a remarkable ability to assess similarity between a variety of stimuli and abstract concepts. Speculations about the preeminence of this feature are common in cognitive science and psychology. Given that our semantic similarity measure for OSM, NetLexSiM, constitutes a core contribution of this thesis, we devote particular attention to this broad area of research.

This section starts by surveying the key ideas in similarity research, and then focuses on the difference between semantic similarity and relatedness (Section 2.5.1). Related work on semantic similarity in the geospatial domain is discussed (Section 2.5.2). As the first component of our similarity measure NetLexSiM, we include network-based, graph-theoretical measures such as P-Rank and SimRank (Section 2.5.3). The second component is lexical similarity, which relies on natural language processing techniques such as WordNet-based similarity measures and paraphrase detection (Section 2.5.4). To evaluate similarity measures, a popular approach is grounded on the comparison with human-generated 'gold standards' (Section 2.5.5). Finally, we discuss related work relevant to the idea of a holistic, viewport-based similarity approach (Section 2.5.6).

The importance of similarity has been discussed first in relation to the workings of human memory in antiquity. In his discussion of theories of memory and recollection, Aristotle [13] highlighted the role of temporal contiguity, contrast and similarity in how the mind associates memories. In the 18th century, Scottish empiricist David Hume [128] developed a theory of association of ideas based on three principles: space-time contiguity, causal relations, and resemblance, i.e. similarity. Austrian philosopher Ludwig Wittgenstein [309] remarked that the meaning of words flows through "a complicated network of similarities overlapping and criss-crossing," rejecting the idea that concepts can be given clear and definitive boundaries (par. 66). Such a 'complicated

network of similarities' seems to occupy a central position in human language and reasoning.

More recently, similarity has attracted attention in the context of cognitive science. Vosniadou and Ortony [300] have stated that the "ability to perceive similarities and analogies is one of the most fundamental aspects of human cognition" (p. 1). Defining his theory of conceptual spaces, Gärdenfors [88] argued that "judgments of similarity are central for a large number of cognitive processes. Judgments of similarity reveal the dimensions of our perceptions and their structures" (p. 5). Furthermore, Goldstone and Son [95] confirm that "assessments of similarity are fundamental to cognition because similarities in the world are revealing. The world is an orderly enough place that similar objects and events tend to behave similarly" (p. 13).

In modern psychology, several similarity theories have emerged to account for a wide range of perceptual phenomena. Goldstone and Son [95] provide a recent survey on cognitive similarity theories. As discussed in relation to semantics, the *context* in which a similarity judgement is performed bears great importance. In particular, similarity is only meaningful when defined with *respect* to something [100]. Different contexts can result in different similarity judgements. Following Goldstone and Son [95], the most important approaches to understand how the mind assesses the similarity of given stimuli can be classified as follows:

Geometric models. These models are the most popular in the analysis of similarity. In a seminal work, Roger N. Shepard [266] has advocated multidimensional scaling (MDS) as a general model for perceptual phenomena. A stimulus is represented as a point or a region in a multidimensional vector space, modelling relevant aspects through continuous variables. Shepard argued that this model seems to match that of biological organisms. Gärdenfors [88] extends this approach by including not only perceptual stimuli, but abstract concepts as well. MDS models have been widely used in computer science, for example to compute the similarity of documents in IR.

Featural models. Amos Tversky [295] has critised MDS models, pointing out that similarity seems to violate geometric properties such as symmetry, minimality, and triangular inequality. To overcome these issues, he developed a seminal psychological theory of similarity based on 'features.' Similarity is computed through a set-theoretical ratio model that weights the common and distinctive features of two stimuli. Although featural models differ from geometric models because they handle discrete instead of continuous variables, they cannot handle structured concepts.

Alignment and transformational models. According to some researchers, representations of concepts are not manipulated by the mind as multidimensional vectors or sets of features, but as hierarchical structures [94, 196]. These approaches are commonly used in computer science when knowledge is represented in ontologies. The computation of similarity between sub-graphs can be based on the effort required for the

alignmentment of the two structures. Two structures are similar to the degree they are easy to align [95]. In an analogous way, similarity can be expressed in relation to the number and nature of transformations required to change one object into another.

As this brief survey has shown, several alternative approaches exist to model the psychological assessment of similarity. This research area shows a remarkably high degree of disagreement, to the point where the term 'similarity' is deemed to be meaningless without further qualifications, making its expressive power and generalisability highly debatable [95]. However, for the purpose of this thesis, it is important to remember the different preoccupations at the centre of psychological and computer science research. While it might well be the case that a universal cognitive theory of similarity is impossible to formulate, computational techniques can reach satisfactory results in many applications, placidly ignoring the difficulties of their theoretical ground. These differences notwithstanding, advances in similarity research can arise from the interplay between cognitive psychology and computer science.

As computer-based techniques aim at emulating aspects of complex human behaviour, similarity is a ubiquitous concept. Clustering, information retrieval, pattern recognition, data mining, image analysis, and recommender systems rely heavily on some measure of similarity between text documents, images, vectors, concepts, and other digital objects [303, 151, 182]. The next section discusses in detail the existing definitions of semantic relatedness and similarity.

2.5.1 Semantic relatedness and similarity

This section discusses two concepts that are at the centre of this thesis, namely *semantic relatedness* and *semantic similarity*, in the context of the geographic information. In the literature on computational semantics, several terms are used inconsistently, including semantic relatedness, relational similarity, taxonomical similarity, semantic association, analogy, and attributional similarity [293]. These terms are often used interchangeably [39]. A prominent example of this tendency is the paper title 'WordNet::Similarity: measuring the relatedness of concepts' by Pedersen et al. [234].

Broadly speaking, psychological similarity focuses on the assessment between perceptual stimuli (visual, aural, etc), while semantic similarity aims at capturing the difference in meaning in markers – signs, to use the semiotic terminology – such as words, sentences, schematic drawings, etc. (see Section 2.4 for a review on semantic theories). Our work focuses on geographic concepts, and the markers used to refer to them, i.e. OSM tags.[13] To relate this to the WordNet terminology, 'words' are semantic markers, 'word senses' refer to concepts.

In the context of semantic networks, Rada et al. [239] suggested that semantic relatedness is "based on an aggregate of the interconnections between

[13]For this reason, we use the terms 'OSM concept' and 'OSM tag' interchangeably.

the concepts" (p. 18). To obtain semantic similarity, the observation must be restricted to taxonomic *is_a* relationships between concepts. Philip Resnik [245] followed this approach, and defined semantic similarity and relatedness as follows: "Semantic similarity represents a special case of semantic relatedness: for example, cars and gasoline would seem to be more closely related than, say, cars and bicycles, but the latter pair are certainly more similar" (p. 448).

More recently, Peter D. Turney [293] added a further distinction between 'attributional' and 'relational similarity.' Following the approach outlined by Medin et al. [201], 'attributes' are statements about a concept that take only one parameter, e.g. $X\ is\ red$, $X\ is\ long$. Therefore, attributional similarity measures the correspondence between the attributes of the two concepts. 'Relations,' on the other hand, are statements that take two or more parameter, e.g. $X\ is\ a\ Y, X\ is\ longer\ than\ Y$. Hence, relational similarity is based on the common relations between two pairs of concepts [293]. On these assumptions, synonymy is seen as a high degree of attributional similarity between two concepts, e.g. river,stream. Analogy, by contrast, is characterised as a high degree of relational similarity between two pairs of concepts, e.g. boat,river and car,road.

Several terms are used in the literature to refer to semantic similarity and relatedness. 'Semantic distance,' for example, is used to refer to the distance between two concepts represented in a geometric semantic model, i.e. vector space model, or in a semantic network, i.e. shortest path [83, 38]. Depending on what attributes and relations are considered, semantic similarity and relatedness can be computed as inversely proportional to semantic distance. Furthermore, the term 'semantic association' is used to define semantic relatedness, in particular in human memory retrieval processes [106, 107]. 'Taxonomical similarity,' on the other hand, is equivalent to semantic similarity [293, 190]. 'Semantic dissimilarity' is often used as a simple counterpart of similarity, but, as Janowicz et al. [135] pointed out, more caution is needed.

A useful concept is that of Lehrer's semantic fields [174, 175]. A 'semantic field' is a set of concepts that describe a restricted domain of human knowledge. The semantic field for 'transport', for example, contains the concepts 'road', 'vehicle', 'fuel', 'motorway', 'accident', 'traffic', and so on. Concepts belong to the same semantic field when they are connected by semantic relations. Common 'semantic relations' are synonymy (A coincides with B), antonymy (A is the opposite of B), hyponymy (A is a B), hypernymy (B is a A), holonymy (B is part of A), meronymy (A is part of B), causality (A causes B), temporal contiguity (A occurs at the same time as B), and function (A is used to perform B).

Khoo and Na [153] have surveyed these semantic relations, whilst Morris and Hirst [214] have explored other non-classical semantic relations. As Khoo and Na [153] remarked, semantic relations are characterised by productivity (new relations can be easily created), uncountability (semantic relations are an open class and cannot be counted), and predictability (they follow general, recurring patterns). In the geographic domain, spatial relations such as proximity (A is near B), and containment (A is within B) have an impact on se-

mantics [263]. Our definitions of geo-semantic similarity and relatedness are outlined in Section 3.5. The next section surveys related work in the area of semantic similarity for the geospatial domain.

2.5.2 Geo-semantic similarity

This section discusses approaches to the computation of similarity that have emerged from GIScience over the past decade, framing them in the context of our work on semantic similarity for OSM concepts. Our contribution to this area consists of the measure NetLexSiM (see Sections 3.6 and 3.7).

Several semantic similarity measures for geo-data have been devised at the intersection between cognitive science, psychology, ontology engineering, GIScience, and IR [134, 146, 179]. A detailed survey of geo-similarity has recently been conducted by Schwering [261], including geometric, featural, network, alignment, and transformational models. Notably, Rodríguez and Egenhofer [250] have extended Tversky's ratio model in their Matching-Distance Similarity Measure (MDSM), one of the first measures of semantic similarity tailored to geographic concepts. MDSM takes context explicitly into account, by selecting a subset of features based on user needs. The MDSM similarity function $S(c_1, c_2)$ is a linear combination of similarity values for parts (p), functions (f), and attributes (a), where ω are the weights and c_1 and c_2 are geographic concepts:

$$S(c_1, c_2) = \omega_p \cdot S_p(c_1, c_2) + \omega_f \cdot S_f(c_1, c_2) + \omega_a \cdot S_a(c_1, c_2) \qquad (2.1)$$
$$\omega_p + \omega_f + \omega_a = 1$$

The specific similarity scores S_p,S_f, and S_a are computed using Tversky's ratio model, which is derived from the ratio between different and common features [295]. This measure takes *context* into account, allowing the user to select a sub-domain of the entire ontology, and assigns contextual weights ω following two strategies. The first is a weighting mechanism based on *variability* that assigns the relevance of a feature based on the feature's informativeness, i.e. low if a feature is shared by all entity classes of the domain, high if it is very specific. The second strategy, called *commonality*, associates the relevance of features with the feature's contribution to the characterization of the application context. For example, if the user selects 'play a sport' as a context, the features containing 'play' and 'sport' obtain higher weights.

All its merits notwithstanding, MDSM relies on a particular conceptualisation, in which each geographic concept has features modelled as *parts*, *attributes*, and *functions*. The conceptualisation is richer and, indeed, more complicated than the semantic model of OSM, on which we focus. The OSM concepts are loosely structured, do not contain explicit features, and therefore cannot be compared using a featural model such as MDSM.

Janowicz et al. [132] have developed Sim-DL, a similarity measure for geographic concepts based on description logic (DL), a popular Semantic Web

language [16]. This approach aims at combining recent psychological theories of similarity with a formalism commonly used in AI and, more recently, in the Semantic Web. Sim-DL compares primitive concepts, roles, and cardinality restrictions, and computes similarity as a weighted sum with respect to these properties. An evaluation of its cognitive plausibility has been carried out on a very small set of concepts [133].

Along similar lines, Bakillah et al. [17] have developed Sim-Net, an approach to computing similarity for *ad hoc* networks, using DL to support concept mapping between ontologies. Although Sim-DL and Sim-Net are promising approaches in some respects, they can only be applied to a knowledge based expressed in a specific Semantic Web/AI formalism such as DL. As observed in relation to MDSM, OSM concepts are described through an informal semantic model, making Sim-DL impossible to apply.

Schwering and Raubal [263] proposed a technique to include spatial relations in the computation of concept-to-concept similarity. For example, the concept 'flooding area' is spatially connected to other concepts, such as 'river', 'riverbank', 'sea', etc., and these relationships can be exploited to increase precision in the similarity computation. They also devised a geometric similarity measure for concepts modelled not as simple multidimensional points but as regions, based on Gärdenfors's conceptual spaces [262].

Keßler [150] surveys the idea of *context* in existing geo-similarity measures. All approaches share the idea of adjusting the weights of the dimensions that are considered in the similarity measurement, to reflect a specificity of the context. The context is defined by restricting the computation to a subset of terms/objects contained in the knowledge base. The importance of context was re-stated by Janowicz et al. [135], who have developed a framework to clarify the semantics of similarity, offering a formal ground to dispel the ambiguity observable in this research area.

All the aforementioned measures focus on classes of geographic entities, and not instances (e.g. *city*, not *Dublin*). Moreover, these measures are knowledge-based, and top-down. Geographic concepts have to be manually constructed in a knowledge base, expressing them in particular formalisms, such as Web Ontology Language (OWL), or description logic (DL). However, in the context of VGI, such knowledge bases might be incomplete, noisy, fragmented, ambiguous, or absent altogether. A suitable approach to similarity in these contexts may be bottom-up, and based on data mining, extracting similarity directly from concept instances. Initial work in this direction has been done by Mülligann et al. [215], who propose a bottom-up semantic similarity measure for OSM based on the change history in instances.

Ahlqvist [2] has extended conceptual spaces to handle semantic uncertainty, using fuzzy sets. As noise and imprecision affect all information, it is reasonable to include them into a fuzzy conceptualisation of geographic entities, allowing for fuzzy properties. For example, a forest is not perceived as being 'closed' or 'open', but has a range of intermediate states. Ahlqvist [3] has also investigated the usage of semantic similarity to capture semantic category change in landscape. Recently, he has suggested that geometric and featural

models of semantic similarity are not radically different as is normally implied in the literature [4]. The next section focuses on the related work for the first component of NetLexSiM, the network-based similarity.

2.5.3 Network semantic similarity

This section describes related work relevant to the area of network-based semantic similarity, i.e. techniques to compute similarity of vertices in graphs. This area is of particular interest to our work, because our OSM Semantic Network, extracted from the OSM Wiki website, encodes geographic concepts in a crowdsourced, open graph, and contributes to our similarity measure NetLexSiM (see Section 3.3).

Semantic networks are a well established approach to knowledge representation, originally devised in psychology to explore semantic theories, and then adopted in computer science, cognitive science, and AI [52, 239, 274]. The vertices of the network represent concepts, and the edges model semantic relationships that connect the edges. According to the classification proposed by Schwering [261], network models are used to measure similarity in semantic networks. These approaches to similarity are based on some form of structural distance between nodes (e.g. edge counting), sometimes adding additional parameters to weight the paths [245], random walks in graphs [240], or on the topological comparison of subgraphs [181]. Such network-based techniques generally rely on well-defined, expert-generated semantic networks such as WordNet [74]. Section 2.5.4 surveys techniques specifically tailored on WordNet.

However, many real-world large collections of data on the Internet do not present such a carefully controlled structure, but encode valuable information in the form of graphs of inter-linked objects. General hyper-links, for example, indicate some relationship between the source and the target page, without specifying their semantics. Given the spread of such networks in many fields, several algorithms have emerged to identify similar objects exclusively on their link patterns in a network that does not explicitly encode attributes, parts, and other details of concepts.

In 1973, Small [271] published the 'co-citation' algorithm. Given a directed graph representing scientific papers and their mutual citations, co-citation measures the similarity between two given papers by the frequency in which they are cited together. Extending co-citation to an iterative form, Jeh and Widom [136] in 2002 created SimRank, a graph-theoretical approach to calculating vertex similarity in directed graphs. The underlying recursive assumption is that "two objects are similar if they are referenced by similar objects" (p. 541). Given its generality and effectiveness, SimRank has attracted notable research interest [183, 178].

The P-Rank algorithm (Penetrating Rank) generalises SimRank, taking into account outgoing links, stating that "two entities are similar if (1) they are referenced by similar entities; and (2) they reference similar entities" [318, p. 553]. Classic algorithms such as the original Co-citation [271], Coupling

[152], and Amsler [9] are specific cases of P-Rank. As recent surveys within GIScience do not address these approaches, the community does not seem to have explored their potential to assess semantic similarity within the geographic domain, favouring other models [261, 135].

These algorithms are included in our approach to semantic similarity for OSM concepts, NetLexSiM, outlined in Section 3.6. The co-citation approach was then applied to our OSM Semantic Network, and is evaluated in Sections 5.5 and 6.3.

2.5.4 Lexical semantic similarity

On the OSM Wiki website, OSM concepts are described through crowdsourced lexical definitions. To obtain a cognitively plausible measure of semantic similarity, such lexical definitions are certainly useful. This section surveys related work in the area of semantic similarity in computational linguistics, which informs the lexical component of our NetLexSiM.

Semantic similarity and relatedness can be computed between two terms in isolation, or between two chunks of text [170, 180]. This section describes term-to-term similarity measures, WordNet-based measures, text-to-text measures, and finally reviews existing evaluations of their cognitive plausibility. To compute the semantic similarity of two geographic concepts in OSM, we compute the semantic similarity of their definitions, i.e. short sections of text, combining all aforementioned aspects of lexical similarity. This vast body of research strongly informs our lexical similarity measure between two concepts, outlined in Section 3.7, and evaluated in Sections 5.6 and 6.4.

Term-to-term semantic similarity. A semantic similarity measure quantifies the association between two given terms ($sim(t_a, t_b) \in \Re$). For example, terms such as 'river' and 'stream' are expected to have higher similarity than 'river' and 'school.' Computing this type of similarity has countless applications in natural language processing, information retrieval, classification, sentiment analysis, ontology alignment, and GIR [238]. Approaches to compute semantic similarity of individual words (as opposed to larger semantic entities) can be classified in two main families: *knowledge-based* and *corpus-based*. Hybrid approaches have also been devised [1].

Knowledge-based (or ontology-based) techniques rely on representational artifacts, such as semantic networks, taxonomies, folksonomies, or full-fledged ontologies [205, 108]. Under a structuralist assumption, most of these techniques observe the relationships that link the terms, assuming for example that the ontological distance is inversely proportional to the semantic similarity [239]. The lexical database WordNet,[14] because of its focus on semantic relationships and dense connectivity, has been successfully used as a support tool to compute similarity [74]. Several measures have been devised and tested to compute generic lexical relatedness and similarity on WordNet (see Sec-

[14] `http://wordnet.princeton.edu`

tion 2.5.4) [170, 137, 313, 124]. EuroWordNet extends the American WordNet network to several European languages [301].[15] These knowledge bases were populated manually, without machine learning, to ensure high quality at the cost of a substantial amount of labour and a relatively low coverage. With the emergence of Web 2.0 and crowdsourcing, larger semantic networks have become freely available (see Section 2.3.2 for a survey of open geo-knowledge bases). In addition, Wikipedia has been transformed into a knowledge base [299, 14], and can be used to compute semantic similarity and relatedness [279, 314].

Corpus-based techniques, on the other hand, do not need explicit relationships between terms, and compute semantic similarity of two terms based on their co-occurrence in a large corpus of text documents. An underlying assumption to these techniques was famously put forward by Harris [117] as the 'distributional hypothesis,' which states that words that occur in the same contexts tend to have similar meanings. Furthermore, Firth [80] suggested that "[y]ou shall know a word by the company it keeps" (p. 11). Simple set-theoretical similarity measures, such as the Jaccard Index, Dice, Tanimoto, and Tversky, rely solely on the number of occurrences of terms in the corpus [295]. Turney's pointwise mutual information (PMI-IR) also uses the occurrences of terms in the Web as a source of semantic similarity of terms [292]. More recently, Sánchez et al. [256] combined knowledge-based techniques with the information content of terms inferred from Web searches.

More complex similarity techniques have been devised, going beyond the simple co-occurrence measures to uncover less visible patterns. The Hyperspace Analog to Language (HAL) model derives a high-dimensional semantic representations from lexical co-occurrence [41]. On similar ground, Latent Semantic Analysis (LSA) has become a prominent approach to extracting a similarity model from a text corpus [167]. The underlying semantic assumption of LSA has been defined by Landauer et al. [166] as follows: "the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and set of words to each other" (p. 259). To achieve this, LSA constructs a large term co-occurrence matrix, reduces it to a manageable number of dimensions with Singular Value Decomposition (SVD), and computes similarity between two terms using cosine distance. Instead of only looking at the total number of co-occurrences, LSA considers detailed patterns of co-occurrence in individual sentences.

Overall, LSA has turned out to be highly effective for a range of problems in natural language processing, but until recently its computational complexity has made its usage problematic on very large corpora [167]. Despite its initial success, HAL has turned out to be less effective than LSA in computing semantic similarity of text [130].

WordNet-based term similarity measures. This section reviews semantic similarity measures tailored to the lexical network WordNet [74]. These

[15] `http://www.illc.uva.nl/EuroWordNet`

measures were included in our approach to lexical similarity, the second component of NetLexSiM (see Section 3.7). To clarify the workings of these measures, it is useful to briefly describe the structure of WordNet. WordNet is a lexical database, containing English words, classified by part-of-speech (e.g. verbs, nouns, etc.). From a structuralist viewpoint, WordNet distinguishes between signifiers (i.e. words, symbols), and signified (i.e. concepts, ideas).

In the WordNet terminology, a word with multiple meanings has several 'word senses.' For example, the word 'field' in WordNet has 17 separate senses, ranging from "a piece of land cleared of trees and usually enclosed" (indicated as *field#n#1*, where n stands for 'noun') to "all of the horses in a particular horse race" (*field#n#12*). A popular application of semantic similarity is the automatic identification of word senses in raw text, called Word Sense Disambiguation (WSD) [233, 28]. Words referring to the same concept are grouped in a 'synset' (a set of synonyms), e.g. *airfield#n#1, landing field#n#1, flying field#n#1*, and *field#n#17* all belong to the same synset. On the other hand, *field#n#1* and *field#n#12* belong to different synsets. It is synsets – and not words – that are linked by semantic relationships, such as meronymity, hypernymity, and hyponymity. To date, WordNet contains 155,287 words grouped in 117,659 synsets.[16]

WordNet-based similarity measures refer to specific synsets, and therefore word senses (e.g. *field#n#3*, *meadow#n#1*). When the word senses are unknown, a common strategy consists of computing the similarity scores between all the word senses, and selecting the highest score. For example, the similarity between *field#n#1* and *meadow#n#1* is different from that between *field#n#9* and *meadow#n#1* [234]. This approach performs reasonably well in general, but fails by overestimating similarity between unusual word senses. As Pantel and Lin [229] pointed out, in WordNet the second sense of the word 'computer' is defined as a person who is an 'expert at calculation.' Without specifying the word senses, the similarity between *computer* and *person* is likely to be largely exaggerated. However, we deem that in our domain this problem affects only a statistically marginal number of cases, and we leave the solution to this problem to future work.

The characteristics of the ten WordNet-based measures are summarised in Table 2.2. This table shows the core ideas of each measure, and whether they rely on network topology, information content, and word sense glosses. For each measure $sim(t_a, t_b)$, a discussion of the difference between similarity and relatedness is included. Details about these measures are reported in Appendix B. Other graph-based techniques have been used on WordNet, in particular inspired by the PageRank algorithm [228]. Hughes and Ramage [127] treated WordNet as a Markov chain and used PageRank, obtaining higher correlation with human datasets. Agirre et al. [1] have developed another version of PageRank on WordNet, and combined it with a corpus-based technique, obtaining high cognitive plausibility at the cost of higher complexity. Along similar lines, Yeh et al. [315] also used PageRank and its random-walk model to compute semantic relatedness exploiting Wikipedia hyperlinks.

[16] `http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`

Name	Reference	Description	SPath	Gloss	InfoC
path	Rada et al. [239]	Edge count	√		
lch	Leacock and Chodorow [170]	Edge count scaled by depth	√		
res	Resnik [245]	Information content of lcs	√		√
jcn	Jiang and Conrath [137]	Information content of lcs and terms	√		√
lin	Lin [180]	Ratio of information content of lcs and terms	√		√
wup	Wu and Palmer [313]	Edge count between lcs and terms	√		
hso	Hirst and St-Onge [124]	Paths in lexical chains	√		
lesk	Banerjee and Pedersen [28]	Extended gloss overlap		√	
vector	Patwardhan and Pedersen [232]	Second order co-occurrence vectors		√	
vectorp	Patwardhan and Pedersen [232]	Pairwise second order co-occurrence vectors		√	

Table 2.2: WordNet-based similarity measures. *SPath*: the measure uses the shortest path in the WordNet taxonomy; *Gloss*: the measure exploits lexical definitions (glosses); *InfoC*: the measure uses the information content of terms. Details in Appendix B.

Text-to-text semantic similarity. This section describes related work in the area of text similarity, i.e. approaches to measure the similarity of two sections of text a and b. This problem has several applications in information retrieval, and natural language processing. Most semantic similarity techniques focus either on individual terms, or on entire documents. However, the definitions of geographic concepts found in the OSM Semantic Network are short texts (the mean length being 48 terms per definition). For this reason we need to focus on sentence, or paragraph similarity.

A classic area of research whose core preoccupation is text-to-text similarity is the detection of plagiarism in academic tests and publications [230, 236]. Anti-plagiarism techniques rely on the distributional hypothesis to detect suspiciously close citation patterns, and similarities in writing styles across different text documents. More recently, the problem of paraphrase detection has become an active research area. For example, the sentence "The Iraqi Foreign Minister warned of disastrous consequences if Turkey launched an invasion of Iraq" should be classified as a paraphrase of "Iraq has warned that a Turkish incursion would have disastrous results" [75, p. 2]. The challenge lies in the detection of sentences that convey roughly the same meaning through a different lexicon. For this reason, a simple set-theoretical measure of exact lexical

overlap does not perform well.

To overcome this problem, Corley and Mihalcea [54] developed a knowledge-based bag-of-words (BOW) technique to paraphrase detection, which relies on some of the WordNet term-to-term measures discussed in Section 2.5.4. In their approach, the term-to-term similarity is also combined with the term specificity, computed through classic Inverse Document Frequency (IDF) [246]. The algorithms compare each term from a with each term from b, selecting only the pairs with maximum similarity. Fernando and Stevenson [75] have extended Corley and Mihalcea's approach by including not only the pairs of terms having maximum similarity, but all the pairs. To achieve this, a complete term-to-term similarity matrix is constructed through a knowledge-based WordNet measure, and its sum quantifies the text-to-text similarity. Compared with Corley and Mihalcea's approach [54], this similarity matrix approach obtains better results in the area of paraphrase detection.

Mihalcea et al. [204] have developed a hybrid approach to text similarity, combining WordNet-based and corpus-based techniques. In particular, they consider PMI-IR [292], and LSA [166] as corpus-based measures, and six WordNet-based measures (*wup*, *res*, *lch*, *jcn*, *lin*, and *lesk*). In terms of precision, the knowledge-based measures outperform the corpus-based ones, and also the combined approach. Tackling similar issues, Islam and Inkpen [130] combined corpus-based term similarity with string similarity to compute the similarity of short texts. Their algorithm, called Semantic Text Similarity (STS), computes a corpus-based term-to-term semantic similarity measure, and includes string similarity to detect misspelt terms. They favour a corpus-based approach, in particular Second Order Co-occurrence PMI (SOC-PMI), over knowledge-based measures for coverage reasons. STS outperforms by a small margin the approach by Mihalcea et al. [204].

Furthermore, random walks on graphs represent a broad area of research, and a promising approach to network-based similarity (see Section 2.5.3) [184, 136, 318]. Ramage et al. [240] apply the random walk approach to the problem of text similarity. After extracting a graph from WordNet, Markov chains are built for each segment of text, using random walks between the terms to reinforce semantically related terms. The semantic similarity of the segments is then computed by comparing the stationary distributions of the chains, used as 'semantic signatures' of the segments. However, the advantages of this approach for textual entailment are not relevant to the OSM definitions.

Because of their conceptual simplicity and effectiveness, the knowledge-based approaches by Corley and Mihalcea [54] and Fernando and Stevenson [75] were included in our technique to compute vector-to-vector semantic similarity between OSM concepts (see Section 3.7). Given the restricted domain covered by the OSM Wiki website, we favour precision over coverage, adopting a knowledge-based approach over existing corpus-based approaches [204]. However, corpus-based approaches shall be considered for future variants of our approach, in particular to include concepts contained in the non-English OSM. In this regard, the methods proposed by Mihalcea et al. [204] and Islam

and Inkpen [130] are certainly among the most appropriate. Our approach to lexical similarity for OSM concepts NetLexSiM is described in Section 3.7.

Evaluations of lexical similarity measures. This section summarises evaluations conducted to assess the cognitive plausibility of semantic similarity measures. The performance of knowledge and corpus-based similarity measures varies greatly depending on the context in which they are applied. In general, the selection of an appropriate similarity measure incurs a precision-coverage dilemma. Knowledge-based measures show higher precision, but suffer from poor term coverage, while corpus-based techniques are less precise, but enjoy substantially wider coverage. In order to assess the behaviour and cognitive plausibility of these similarity measures, empirical evidence has to be gathered. Several evaluations have been published, mostly in the area of psychology and computational linguistics.

Budanitsky and Hirst [38] have investigated the cognitive plausibility of five WordNet-based measures to detect malapropism, identifying the measure by Jiang and Conrath [137] (*jcn*) as the most plausible. As discussed in Section 3.5, semantic relatedness is more general than similarity. Budanitsky and Hirst [39] have analysed in detail the characteristics of these five measures when dealing with relatedness as opposed to similarity. Again, *jcn* turned out to be more cognitively plausible than *hso*, *lch*, and *res*. According to the evaluation by Patwardhan and Pedersen [232], gloss-based techniques (*lesk*, *vector*, and *vectorp*) outperform path-based measures when measured against human datasets by Rubenstein and Goodenough [252], and Miller and Charles [206].

In the area of purely corpus-based techniques, Turney [292] have compared the performance of LSA and a PMI-based algorithm (PMI-IR) in the identification of synonyms. In this task, PMI-IR performs better than LSA, having also the advantage of being remarkably simpler than LSA. Similarly, Recchia and Jones [241] investigated the impact of corpus size on LSA and PMI, suggesting that simple measures such as PMI can compete with LSA when the datasets are large enough. Recently, Wang and Hirst [304] have explored the impact of the WordNet taxonomy depth on similarity measures.

To the best of our knowledge, no evaluation of lexical similarity measures geared towards a geographic domain has been conducted. To fill this knowledge gap, we have included in NetLexSiM, our approach to semantic similarity, two text-to-text, and ten term-to-term measures. This evaluation is described in Sections 5.6 and 6.4.

2.5.5 Similarity gold standards

Semantic similarity measures can be evaluated against a human-generated set of psychological judgements. This section gives an overview of published similarity and relatedness gold standards, mostly from psychology and computational linguistics. Our geographic similarity survey, outlined in Section 6.2, is inscribed in this line of research.

The term 'gold standard,' originally defining a monetary system utilised

in international trade until the 1930s, is described by the Oxford Dictionary of English as "a thing of superior quality which serves as a point of reference against which other things of its type may be compared."[17] In computer science, the term is generally used to describe high-quality, human-generated datasets, capturing human behaviour in relation to a well-defined task. Such datasets can then be used to assess the performance of automatic approaches, by quantifying the similarity between the machine and the human-generated data.

Gold standard-based evaluation is common in natural language processing tasks, such as part-of-speech (POS) tagging, entity resolution, and word sense disambiguation [260, 290, 47, 233]. Adopting this approach, a technique or a model can be deemed to be more or less plausible by observing its correlation with human-generated results. Such datasets are created by combining the results from a number of human subjects who perform a given task, either under controlled conditions, or through online forms. To be considered valid by a research community, a gold standard needs to meet certain criteria, such as coverage, quality, precision, and inter-subject agreement. Disagreements about the validity of a gold standard are quite common and, when weaknesses are uncovered, a gold standard can be demoted to a 'golden calf' [143].

In the area of semantic similarity, several datasets have been published to evaluate the cognitive plausibility of computational techniques to quantify the similarity of two terms (e.g. 'river' and 'stream'). A set of word pairs is ranked or rated by human subjects, and the mean score is considered as a representative score for a word pair. A similarity algorithm is then compared against such a gold standard through Spearman's ρ [276] or Kendall's τ [148]. High correlation, when statistically significant, is interpreted as a sign of high cognitive plausibility (see Section 5.3).

Over the past 50 years, several authors investigating semantic issues in psychology, linguistics, and computer science created similarity datasets. The first similarity gold standard was published in 1965, in a paper in which Rubenstein and Goodenough [252] collected a set of 65 word pairs ranked by their synonymy. Following a similar line of research, Miller and Charles [206] published a similar dataset with 30 word pairs in 1991. More recently, Finkelstein et al. [78] created the WordSimilarity-353 dataset, which contains 353 word pairs ranked by relatedness.[18] The dataset was subsequently extended to distinguish between similarity and relatedness [1].[19] In a study of the retrieval mechanism of memories, Nelson et al. [221] collected associative similarity ratings for 1,016 word pairs.

Given that our work is focused on geographic concepts, it is important to review similarity datasets in the areas of GIS and GIR. In this area, Janowicz et al. [133] conducted a study on the cognitive plausibility of their Sim-DL similarity measure. However, the study was conducted in German on a very small set of concepts, and for this reason it is difficult to reuse in different

[17] http://oxforddictionaries.com/definition/gold%2Bstandard
[18] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353
[19] http://alfonseca.org/eng/research/wordsim353.html

Dataset	Subjects	Task	Objects	Notes
Rubenstein and Goodenough [252]	51 paid college undergrads: group I (15 subjects), group II (36 subjects).	"Order the pairs according to amount of *similarity of meaning*"	65 pairs of 'themes' (ordinary English words), ranging from highly synonymous pairs to semantically unrelated pairs.	No geographic concepts.
Miller and Charles [206]	38 undergraduate students. US English native speakers.	"The subjects were told to judge similarity of meaning." 32 word pairs: 5 Likert scale (0: unrelated, 4: perfect synonymy).	A subset of 30 noun pairs from Rubenstein and Goodenough [252]; 10 high similarity pairs; 10 intermediate similarity pairs; and 10 low similarity pairs.	No geographic concepts.
WordSimilarity-353, Finkelstein et al. [78], Agirre et al. [1]	13 experts for first set (153 pairs), 16 experts for second set (200 pairs). Near-native English proficiency.	"[E]stimate the *relatedness* (or *similarity*) of the words in pairs on a scale from 0 (totally unrelated words) to 10 (very much related or identical words)"	353 pairs of manually selected nouns and compound nouns.	No geographic concepts. It considers semantic similarity and relatedness.
MDSM evaluation dataset, Rodríguez and Egenhofer [250]	72 paid undergrad students (two groups of 36 people). US English native speakers.	"[R]ank places according to your judgment of similarity." 2 surveys (A and B): 5 questions each, 10-11 pairs to rank in a given context.	Concepts from WordNet and SDTS. Concepts have attributes, parts, and functions. 33 geo-concepts.	Focused on geographic concepts, but low coverage.
Nelson et al. [221]	94 undergraduate students rewarded with academic credits	"[R]ate each pair of words that you are given on a scale of 1-7 in terms of their degree of mutual association or relatedness."	1,016 word pairs selected unsystematically from a cued recall database of 2,000+ pairs.	No geographic concepts. Largest dataset to date. Focused on general semantic relatedness and not similarity.
Janowicz et al. [133]	28 unpaid subjects (20-30 years of age).	"[A]ssess the similarity between the description of the search concept and every description of the target concepts." Continuous similarity scale and Likert scale to assess confidence level (from *not sure* to *sure*).	Six geographic concepts related to bodies of water.	In German, not applicable to English concept definitions. Low coverage: only 6 concepts.

Table 2.3: Survey of lexical similarity gold standards.

contexts. In order to evaluate their MDSM, Rodríguez and Egenhofer [250] collected similarity judgements for 33 geographic concepts, including large natural entities (e.g. *mountain* and *forest*), and man-made features (e.g. *bridge* and *house*). To date, the MDSM evaluation dataset is certainly the most significant similarity gold standard for geographic concepts. For this reason, this dataset was utilised to carry out a preliminary evaluation of our network and lexical similarity measures (see Section 3.6 and 3.7). Section 5.4 describes the MDSM evaluation dataset in detail.

The salient characteristics of these gold standards are summarised in Table 2.3, detailing the subjects, the assigned task, and the objects evaluated in each dataset. It is important to note that most authors did not have the explicit intention to construct a similarity gold standard, but rather to analyse a specific aspect of semantic similarity or relatedness. In some cases, these datasets were treated as gold standards in the subsequent literature [252, 206]. To our knowledge, only the WordSimilarity-353 was explicitly designed to be a similarity gold standard [78, 1]. Some of these gold standards have been extensively utilised to assess general term-to-term similarity measures [252, 206, 78]. In the geographic context, only the MDSM evaluation dataset is suitable to evaluate semantic similarity of geographic concepts [250]. However, no existing dataset offers a sufficient number of geographic concept pairs, both for semantic similarity and relatedness. Furthermore, some datasets do not distinguish between semantic similarity and relatedness [78, 221].

To overcome these limitations of existing gold similarity standards, we have developed the Geo Relatedness and Similarity Dataset (GeReSiD), following this well-established tradition of research on semantic similarity. This dataset is designed to offer an evaluation testbed for semantic similarity measures, focusing on geographic concepts commonly found in web maps, and distinguishing clearly between semantic *similarity* and semantic *relatedness*. GeReSiD is described in Section 6.2.

2.5.6 Viewport similarity

This section reviews related work relevant to our holistic approach to GIR based on viewport similarity, outlined in Section 3.9, and evaluated in Section 6.6. OSM tags are used to specify the semantic content of a map feature, and measuring a tag-to-tag semantic similarity is certainly the main building block of a similarity measure for volunteered geographic concepts. A strongly related issue is that of the semantic similarity of viewports, which focuses on sets of features as opposed to individual features. Our work in this area has been published in [23].

In web mapping services, geographic data is distributed through a viewport, a rectangular viewing frame that represents a geographic area at a given scale. Typically, users can zoom and pan, updating the viewport. The concept of viewport is inscribed in a long-standing representational tradition of the screen. Manovich [194] traces a compelling genealogy of the screen, seen as a flat, rectangular surface "acting as a window into another space" (p. 115).

While interacting with map viewports, users aim at fulfilling their spatial information need. This process is often focused on specific individual map features, decomposing the represented landscape analytically. However, the field of landscape ecology strongly argues that landscape is perceived holistically, as a complex whole. Antrop states that the holistic approach was stimulated by aerial photography, which represents the landscape in its holistic complexity [12]. Naveh [219], in his broad discussion on landscape ecology and system theory, identifies holism as "perceiving all parts in their full context", and criticises analytical, reductionist approaches "focused on single, isolated parts of the system" (p. 13). Moreover, in cognitive science and psychology, holistic cognition is believed to play a major role in perception [273, 222].

To be interpreted by humans, the geographic information represented in viewports has to convey some intelligible meaning. The semantics of geographic data has been discussed extensively by Kuhn [161], who points out the difficulties of grounding meaning in symbolic systems. It is a tautology to state that meaning is crucial in geographic information retrieval (GIR), which aims at identifying relevant features in large datasets [238]. To date, most GIR systems focus on individual map features, with particular emphasis on text-based retrieval [91].

In order to compare, classify, index and cluster geographic objects by their semantics, several analytical approaches have been devised [250, 133]. In these approaches, the similarity of individual semantic geographic concepts are compared based on their commonalities, differences, positions in taxonomies, and so on (see Section 2.5.2). While such models can compute plausible similarities between specific features or feature types, they do not consider the computation of holistic similarity of the map fragments that are, ultimately, displayed to and manipulated by the users in rectangular viewports. When presented to users, viewports are bi-dimensional images. For this reason, our approach can be seen as analogous to techniques used in image processing systems to compare raster images [284]. However, while such techniques are based on the analysis of low-level image features, such as colour, to compute the similarity of raster viewports, our focus is on the semantics of specific objects. Therefore, our GIR system focuses exclusively on vector data, regardless of the particular visual display of the rendered viewport.

Moreover, viewports show geo-information at a specific map scale. In a typical web map, a viewport is associated with a scale and includes different types of features depending on specific visibility rules. None of the traditional semantic similarity measures discussed above take scale into account, as they focus on abstract psychological classes rather than on viewports. Our system, on the other hand, captures the scale of a viewport by building a semantic descriptor that includes only features that are present (i.e. represented) at the viewport scale – but independently of how they are represented. Our holistic approach to computing semantic similarity of viewports is described in detail in Section 3.9, and evaluated in Section 6.6.

2.6 Summary

This chapter has provided an overview of the research areas related to our work. The research presented in this thesis is located within the growing scientific corpus on spatial crowdsourcing, Volunteered Geographic Information (VGI), geo-semantics, semantic networks, lexical semantics, and semantic similarity measures. We started by reviewing the background to our preliminary work in map personalisation, user profiling, and automatic collection of feedback (Section 2.2). The rise of spatial crowdsourcing, VGI, and OpenStreetMap (OSM) was subsequently discussed in detail, providing a survey of open geo-knowledge bases (Section 2.3). The broad area of semantics was discussed, focusing on logical and linguistic theories, and covering in detail geo-semantics, a growing area of Geographic Information Science (GIScience) (Section 2.4).

The literature about the other core aspect of our research – the similarity of geospatial concepts – was thoroughly investigated (Section 2.5). After addressing the main tenets of psychological and cognitive research on similarity, we surveyed network-based, graph-theoretical similarity measures, and lexical measures in the area of natural language processing, both applied to the Network-Lexical Similarity Measure (NetLexSiM). In the context of GIScience and VGI, semantics and similarity theories provide important foundations to the efforts to tackle semantic issues. Our approach to provide semantic support for OSM are greatly indebted to the vast body of work we have summarised in this chapter.

CHAPTER

THREE

APPROACH

3.1 Overview

Given the growing impact of Volunteered Geographic Information (VGI), the issue of data semantics is paramount for effective exploitation of user-generated geo-datasets. The lack of a formal semantic structure in OpenStreetMap (OSM), the leading grassroots mapping project, hinders the effective usage of its large and evolving data, offering often ambiguous and fragmented meta-data. Our approach to tackling this issue in the context of OSM consists of creating a novel semantic support resource, the OSM Semantic Network, and a new approach to computing semantic similarity for geographic concepts, the Network-Lexical Similarity Measure (NetLexSiM). This chapter details our contribution to providing semantic support for OSM data for a number of advanced applications, including Geographic Information Retrieval (GIR), map personalisation, user profiling, and information integration.

First, we describe the preliminary work in map personalisation in which we noticed the semantic problems of the OSM vector dataset (Section 3.2). Our spatial recommender system, RecoMap, monitors user interaction, and recommends map features based on users' historic behaviour. RecoMap's approach focused on individual OSM features, relying on a basic semantic model. To overcome its limitations, we devised a technique to enrich the OSM dataset with open geo-knowledge bases, integrating rich spatial data with rich semantic data. To provide support for OSM semantics, a novel open source resource was devised, the OSM Semantic Network, freely available online (Section 3.3).[1] This semantic network can be used to compute the semantic similarity of geographic concepts, an endeavour with wide-ranging applications in map personalisation, data mining, and GIR. To clarify the nature and scope of our approach to OSM semantics, we distinguish between 'weak' and 'strong' semantics (Section 3.4).

A specific definition of semantic similarity, contrasted with semantic relatedness, is subsequently provided (Section 3.5). The resulting approach to semantic similarity in the OSM Semantic Network, NetLexSiM, is based on two complementary aspects. First, we outline an approach to network-based semantic similarity, based on co-citation measures (Section 3.6). Second, a

[1] `http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork`

knowledge-based technique to compare lexical definitions of concepts is presented (Section 3.7). The two components of NetLexSiM are integrated in a hybrid similarity measure, which aims at overcoming the limitations of both network and lexical similarities (Section 3.8).

Finally, concept-to-concept is not the only type of semantic similarity that can be computed to facilitate the usage of VGI data. Instead of focusing on individual concepts, a GIR approach can consider viewports as unified semantic units. Our holistic, viewport-based GIR system adopts this perspective, providing a mechanism to extract semantic descriptors for viewports. Such descriptors can be indexed, and searched with a query-by-example interface, enabling an alternative approach to geographic data discovery (Section 3.9).

3.2 Preliminary work

This section reports on the first phase of our contribution to geospatial research. This preliminary work, conducted in the area of map personalisation and implicit feedback, highlighted semantic issues in the OSM dataset, indicating a research direction for VGI semantics that we have followed in this thesis. The remainder of this section outlines the RecoMap prototype, a spatial recommender system (Section 3.2.1). In order to expand the conceptualisation of OSM, we devised a technique to combine OSM vector data with a semantically rich ontology such as DBpedia (Section 3.2.2). The limitations of these endeavours stressed the need for a richer conceptualisation of OSM data, which will be outlined in Section 3.3.

3.2.1 RecoMap

This section describes a spatial recommender system, which has been developed to explore techniques of implicit feedback analysis. The development of this prototype was carried out in the context of the Online Dublin Computer Science Summer School (ODCSSS).[2] This work was important to identify the research problem addressed in this thesis, and was published in [20].

The system, called RecoMap (Recommender Map), has been developed to investigate the usage of two implicit spatial interest indicators, the user location and the mouse clicks on the map. RecoMap combines dynamic and static profiling to provide the user with personalised spatial recommendations through a combination of user context management and user profiling. The approach extends previous work on implicit feedback analysis by considering the user context when making personalised recommendations [189]. The RecoMap prototype renders an interactive map of the University College Dublin campus to the user, monitors their location (acquired from a GPS sensor) and the mouse clicks on the map. The vector map data consists of polygons, polylines and points of interest taken from the OSM vector dataset.

[2] http://www.odcsss.ie

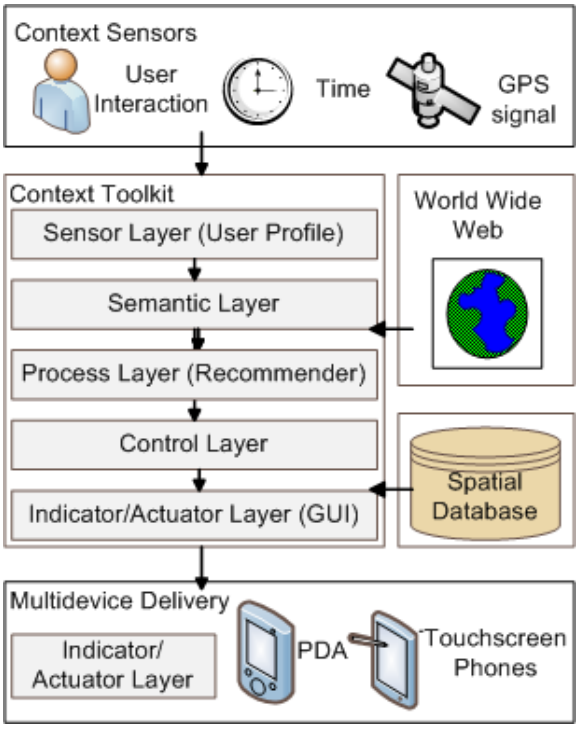

Figure 3.1: RecoMap: system architecture

RecoMap consists of a modular interactive application that renders a vector map and recommends spatial items and associated information by inferring user interests from implicit feedback indicators and user context. The system architecture, shown in Figure 3.1, is based on the context management toolkit proposed by Zimmermann et al. [319]. This toolkit integrates user modeling with context modeling and offers a design architecture very suitable for such applications. The 'context sensors' collect information, and stream it to the Context Management System (CMS), the main component of RecoMap. The CMS receives information about the user context from the sensors and provides recommendations to the user. The Context Toolkit is the core of the CMS and the five layers which constitute it are described below.

The Sensor layer receives information about the user context (user location, an interest radius calculated upon the user speed, time of the day), and their interactions the GUI (mouse clicks and other events). These details are collected in the user profile and are available for the other layers to process. The semantic structure of the system is defined in the Semantic layer. The user world is a set of items stored in a spatial database. An item is either a complex geometry (a polygon) or a point of interest representing a geographical entity which the user might have interest in and interact with. Certain items (such as restaurants, shops, etc.) can be associated with relevant resources on the World Wide Web (such as web sites or web services). These items are displayed on an interactive digital map (Figure 3.2).

The Process layer monitors the evolution of the user profile as it is updated by the Sensor layer. It assigns an interest score to each item located within the current interest radius. The interest score α for the item i is calcu-

lated with Equation 3.1, taking into account both historical interactions (*interaction*) and current user routing distance from the item (*proximity*):

$$\begin{aligned} \alpha_i &= P_i P_R + I_i I_R \\ P_R + I_R = 1, \quad P_i \in [0,1], \quad I_i &\in [0,1] \end{aligned} \tag{3.1}$$

P_R and I_R are respectively the *proximity ratio* and the *interaction ratio*, meaning the weight the system attributes to each indicator. The preliminary evaluation has been conducted with $P_R = 0.2$ and $I_R = 0.8$, emphasising interaction over proximity. Furthermore, the proximity score P_i and the interaction score I_i are normalised between the maximum and minimum score among the items within the proximity score, as illustrated in Equation 3.2:

$$\begin{aligned} P_i &= \frac{|P_i| - |P_{min}|}{|P_{max}| - |P_{min}|}, \quad |P_i| \in [0, \infty) \\ I_i &= \frac{|I_i| - |I_{min}|}{|I_{max}| - |I_{min}|}, \quad |I_i| \in [0, \infty) \end{aligned} \tag{3.2}$$

The variable $|I_i|$ is a non-negative real number incremented by a fixed value at every interaction, representing the degree of interest expressed toward the item i (e.g. a mouse click directly on the item increments $|I_i|$ by 0.01 and a click on a recommended item by 0.05). Similarly, $|P_i|$ is intended to represent the 'spatial interest' shown implicitly by the user position in the physical world: the more the user spends time close to an item, the higher $|P_i|$ becomes. Given the volatile nature of user interests pointed out by Wu et al. [312], a time decay function based upon the days elapsed since the last interaction with the item i is applied on $|I_i|$ and $|P_i|$.

When a certain condition occurs in the user context and profile (e.g. user heads toward a new area), the Control layer triggers an adaptive action. For example, when the user either starts to explore a new area or alters the user profile by interacting with the map, the control layer obtains new recommended items, which have the highest interest score at a given time for a given location. The Indicator/Actuator layer contains the Graphical User Interface (GUI) of RecoMap. The GUI is compatible on multiple platforms including tablet PCs and smartphones. As shown in Figure 3.2, it consists of an interactive map and an information panel. The map has a set of intuitive functions including zoom, pan and drag. The recommended items are visually highlighted with a bold outline and a stronger colour. The current user location and their historical locations from the Sensor layer are presented with a trace and a highlighted icon which helps the user visualise the items' spatial layout.

The information panel, on the right hand side of the GUI, consists of a list of html links and previews of recommended items. When a recommended item is clicked, the panel displays information associated with it and the system acknowledges a successful recommendation by incrementing the interaction score of the item via the Control layer. Below this, the browser window shows further details of the item the user is interacting with. In the system, each geographic item (i) is assigned an interest score (α), which represents the

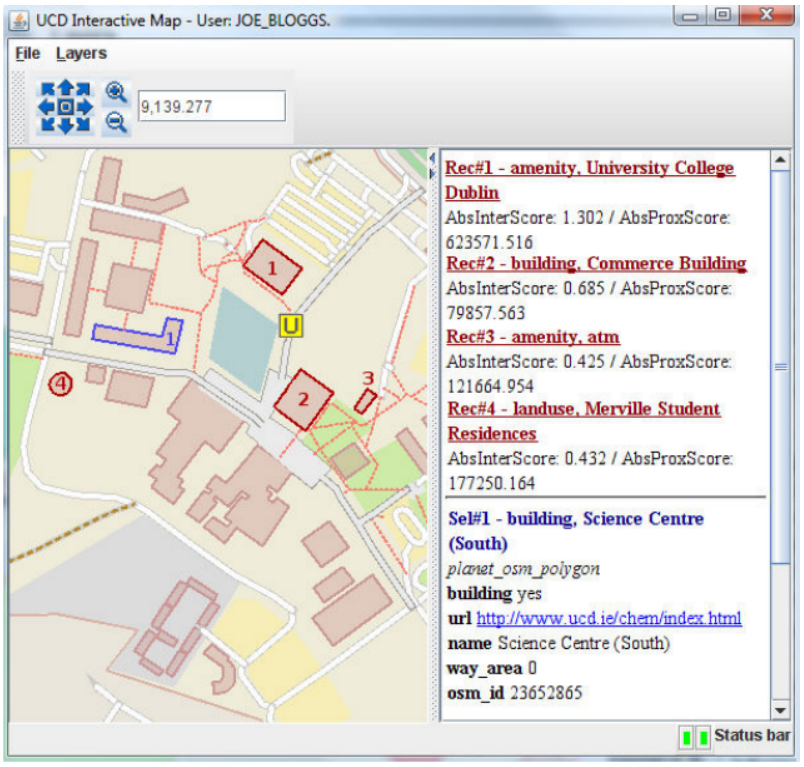

Figure 3.2: RecoMap: map navigation. The U icon shows the user location.

degree of interest that the current user has towards the item. The interest score (α) for the item i is calculated with Equation 4.1, taking into account both historical interactions (*interaction*) and current user routing distance from the item (*proximity*).

$$\begin{aligned} \alpha_i &= P_i P_R + I_i I_R \\ P_R + I_R = 1, \quad P_i \in [0,1], \quad I_i &\in [0,1] \end{aligned} \tag{3.3}$$

where P_R and I_R are respectively the *proximity ratio* and the *interaction ratio*, meaning the weight the system attributes to each indicator.

RecoMap's main contribution involves extending current techniques by incorporating user context to develop novel algorithms and personalise spatial contents. Based on a context management framework, a layered approach is adopted. RecoMap has separate components for sensing the user context and for recording interaction in order to build a dynamic user profile. The technique explored in RecoMap was then implemented in the Web platform for map personalisation and visualisation (see Section 4.4). RecoMap operates at the instance level, i.e. individual spatial features, such as specific buildings and amenities. The OSM features are described using lexical labels, called 'tags' (see Section 2.3.3). Such tags can be used to model user interests, inferring user interests beyond a narrow boolean model.

This work highlighted important limitations in the OSM semantic model. The lack of a formal semantics makes it problematic to infer implicit, latent

aspects of the map features. For example, if a user repeatedly clicks on colleges and universities, it is impossible to infer an interest in 'higher education,' which could refine and improve the user profile, enabling better recommendations and group profiling [283]. The semantic model of OSM, in other words, would benefit from a richer representation of its geographic concepts. As a first step to explore this direction, we have devised a technique to combine OSM features with semantic data from other geo-knowledge bases, outlined in the next section.

3.2.2 Semantic enrichment of OSM

In the OSM vector dataset, geographic features are represented as geometric objects, tagged with lexical labels ('tags' in the OSM jargon). These labels contain little meta-data, and do not provide a rich semantic description of the geographic features. On the other hand, many open geo-knowledge bases, surveyed in Section 2.3.2, host semantically rich feature descriptions with little or no spatial information. This section describes our efforts to bridge the semantic gap between spatial and non-spatial volunteered datasets, in particular DBpedia [34], a Semantic Web-version of Wikipedia, and LinkedGeoData [15], a linked dataset extracted from OSM data. This work has been published in [19].

As a first step to providing a semantic support for OSM, we have developed a system that bridges ontological concepts with geographical entities. The objective of the system is to retrieve semantic content from a geographical location, taking scale into account and finding ontological terms that can be used for user interest extraction. When the user clicks on the Web map, the system processes a spatial query mapping spatial features to semantic entities. This system has been developed as a module of our Web platform for map personalisation and visualisation, developed in collaboration with the National University of Ireland, Maynooth (see Section 4.4). The module is a Web application that processes a spatial query and retrieves ontological results interacting with several Web services.

Spatial discovery queries

Through the web GUI, users can submit spatial discovery queries, whose main parameters are:

- **geo-location**: latitude and longitude (e.g. *53.3071, -6.2218*);
- **radius**: radius expressed either in screen pixels (dependent on the current map scale) or in meters (e.g. *20p* or *500m*);
- **scale**: map scale (e.g. *1/14,000*);
- **max results**: maximum number of OpenStreetMap objects retrieved from CloudMade in the given area (e.g. *10*).

For example, a valid exploratory query could be { *geo-location = (53.3071, -6.2218), radius = (500m), scale = (1/14,000), max results = (10)* }. This query retrieves a maximum of 10 ontological entities located within a radius of 500 metres from the specified point (expressed in the lat/long coordinates 53.3071 and -6.2218) at a low scale (1:14,000, corresponding to the street level). The indicated target geographical area roughly covers the University College Dublin campus. The Semantic Service is then expected to return entities that are semantically related to the target geographical area, such as 'education' as well as individual institutions located in the campus.

In this process, a critical issue is the role of the map scale. Commonly used Web maps (such as Google Maps and Visual Earth) take scale into account for two reasons, firstly to decide which objects need to be displayed, and secondly to choose a suitable visual style aiming to combine clarity and aesthetic coherence. In the context of implicit feedback analysis, scale plays an important role which, as Mac Aoidh et al. [189] pointed out, has not been fully investigated. Given the different visual content of a Web map at different scales, including this parameter in the semantic extraction is beneficial to further reduce the semantic gap.

A complex research question that is worth asking is: what can map scale reveal about the user's spatial interests? For example, when a user chooses a scale of 1:30,000,000, a typical Web map only renders country borders, major lakes and capital cities. If the user operates for a long time over a geographic area at a very high scale, it is reasonable to expect that they are likely to be interested into high regional features such as rivers, urban areas and major infrastructures. Considering all the entities located in the target area, such as individual cinemas and theatres, would be semantically misleading. On the contrary, when the scale is as low as 1:10,000, the user can see objects at the street level such as individual buildings, restaurants and bus stops and therefore the Web map includes all the fine-grained details available in the dataset. A project such as LinkedGeoData does not take this problem into consideration, as it only provides links for cities and few other entities.

Our system follows this idea by retrieving different object categories at different scales, following an approach sometimes called 'what you see is what you get.' The definition of these 'semantic layers' associated with different scales mirrors closely the structure of the CloudMade map, whose internal structure and style are fully customisable. We have chosen to start with the default map style, called 'the original' in the CloudMade style editor[3], because it represents the general-purpose Web maps that have become ubiquitous in neogeography. For instance, the top semantic layer in the scale range (1:28,000,000-1:15,000,000) only includes 'countries' and 'capital cities.' On the contrary the lowest semantic layer, whose scale range is (1:2,000-1:1), includes categories such as restaurants, ATMs and shops, exposing the smallest entities contained in the dataset.

The radius is also an important parameter that can have a high impact on the result of the spatial query. The CloudMade service accepts the query

[3] `http://maps.cloudmade.com/editor`

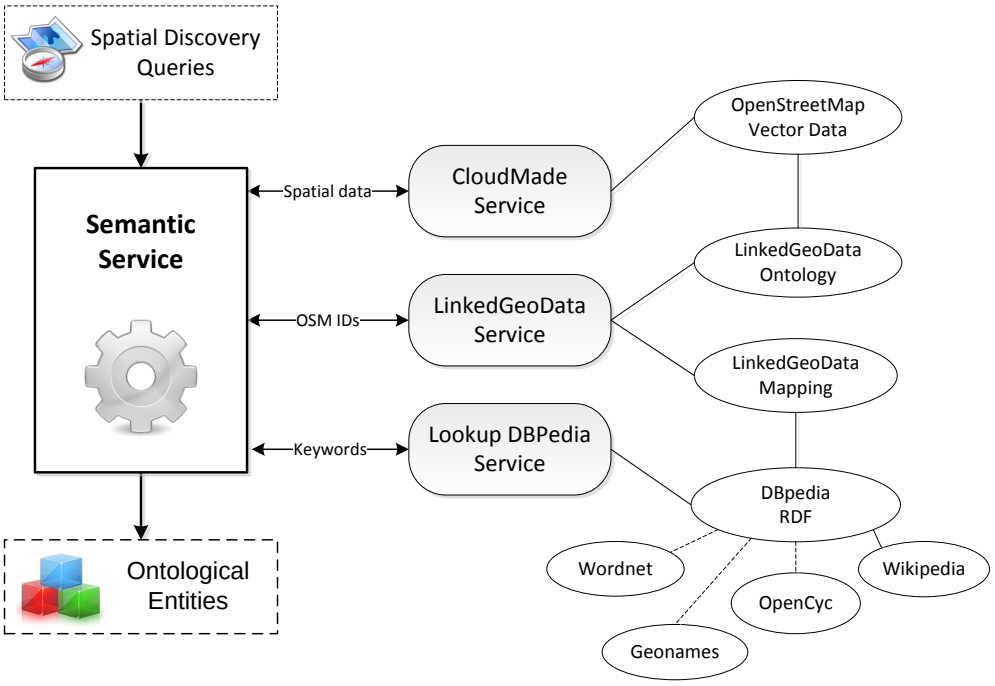

Figure 3.3: Architecture of the Semantic Service

radius in meters. In order to match the user perception more closely, our Semantic Service also accepts the radius in screen pixels, converting them into meters on the map currently being displayed. Another relevant parameter is the maximum number of OpenStreetMap objects to be retrieved. This parameter can be used to tune the query scope, trying to strike a balance between a more inclusive result set potentially including irrelevant entities and a smaller result set, which might exclude relevant ones.

Semantic service

This service hosts the main functionality of the system. The architecture and the interactions of this service are presented in Figure 3.3. The query, including the aforementioned parameters, is processed by the Semantic Service as follows :

1. Retrieve OpenStreetMap objects from CloudMade service.
2. Retrieve DBpedia mapping from LinkedGeoData.
3. Extract key words from OpenStreetMap objects.
4. Lookup DBpedia with extracted key words.
5. Determine heuristically whether the DBpedia nodes are valid or not.
6. Extract ontological terms and categories.
7. Merge results and store them in an XML file.

In **step 1** the system retrieves the OpenStreetMap objects located within a certain radius from the CloudMade service, taking scale into account. The OpenStreetMap nodes contain metadata and tags. For example, the OSM node that represents University College Dublin contains the following XML data:

```
<node id="83211617" lat="53.3071709" lon="-6.2218882"
  user="rorym" uid="23770" visible="true" version="2"
  changeset="432796" timestamp="2008-03-31T00:24:37Z">
    <tag k="amenity" v="university"/>
    <tag k="created_by" v="Potlatch 0.8a"/>
    <tag k="name:en" v="University College Dublin"/>
</node>
```

It is possible to note the nature of semantic information contained in OpenStreetMap. The entity names and the *amenity* tag do not allow further semantic navigation, for example towards the concepts *education*, *school* and *college*, which are highly similar. Such data is unsuitable for implicit feedback analysis, because it does not show connections between those ontological concepts, which are useful to determine user's interests (e.g. in education). Therefore it is necessary to proceed to step 2 to get richer semantics.

In **step 2** the node IDs are then extracted and matched on DBpedia nodes through the LinkedGeoData mapping dataset, described in Section 2.3.2. In the case of the University College Dublin node, the mapped entity on LinkedGeoData[4] does not contain any more information than the original OpenStreetMap node. Afterwards, in **step 3**, key words are extracted from the node, in an effort to obtain useful semantic content. The extraction of key words from OpenStreetMap objects is executed by defining a subset of the tags as semantically relevant, ignoring the others. OpenStreetMap metadata, such as contributor's information and data sources (*created_by*, *user* and *source*) are discarded, as well as tags that do not seem to form points of interest for a general user ('abutters,' 'smoothness,' 'incline,' 'voltage').

Semantically relevant tags are given a high priority in the key words list: the English name tag (*name:en*), when available, has the highest priority, followed by *amenity*, *shop*, *tourism*, *landuse*, *natural*. In the case of the node representing University College Dublin, the extracted keywords are 'university', 'college' and 'dublin', which are utilised in the following step. In order to allow users to retrieve nodes by key words, DBpedia provides a Web service called DBpedia Lookup,[5] designed and used by Kobilarov et al. [156] in collaboration with the BBC. The service takes key words and returns the URI of matching DBpedia nodes, if any.

In **step 4**, the Semantic Service invokes DBpedia lookup and analyses the return URIs. In the example, the service returns the University Dublin College page as a first result.[6] In **step 5** the system utilises a heuristic to determine whether the returned DBpedia node is valid or not, based on two criteria: (a) geographic proximity and (b) tag matching. To assess criterion (a), the system

[4] http://linkedgeodata.org/page/node83211617
[5] http://lookup.dbpedia.org
[6] http://dbpedia.org/page/University_College_Dublin

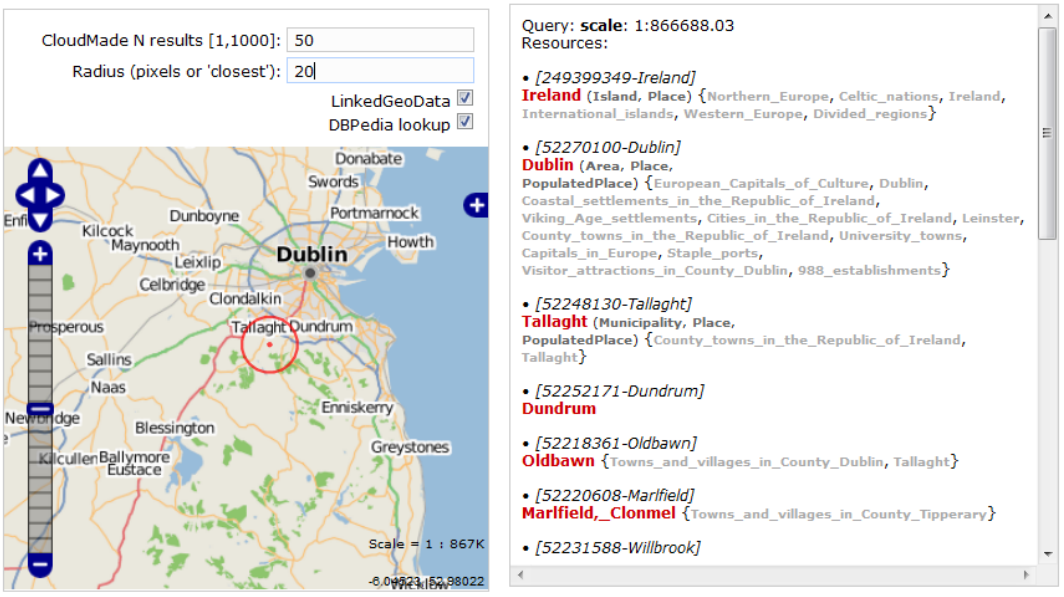

Figure 3.4: Web User Interface for Spatial Queries

calculates the distance between the OSM and the DBpedia node centroids. If the distance is lower than a threshold ϵ, the match is considered valid. A value for ϵ that seems to give reasonably good results is 50km, for example preventing frequent mismatches between European and North-American cities with the same name. In the case of University College Dublin, criterion (a) is fulfilled, with a distance smaller than 1 kilometre. When the geo-location is not available in the DBpedia node, criterion (b) is considered. All the tags present in the OSM node are matched against the DBpedia node. If the matching tags ratio is higher than threshold σ, the node is considered to be valid. The default value of σ, based on a preliminary evaluation, is set to 0.5. The optimal values of ϵ and σ can be determined by further experimental evaluation.

In **step 6**, the retrieved valid nodes are then processed to extract ontological terms and categories, which enhance the semantic relevance of the results. For example, the DBpedia node representing University College Dublin contains, among others, the ontological term 'university.'[7] By visiting this ontological term, it is possible to navigate to the parent term ('educational institution') and, from there, to reach the terms 'school' and 'college.' Semantic similarities starting from the OSM node can now be explored. In **step 7**, all the results are merged and stored in an XML structure and can be either stored for further analysis, or formatted in human-readable HTML code and displayed to a human user.

The approach to information integration is evaluated in Section 5.2. Although this work showed encouraging results in retrieving relevant instances of OSM features in DBpedia, it confirmed the necessity of having a richer representation of the geographic concepts used in the OSM vector dataset. In

[7] `http://dbpedia.org/ontology/University`

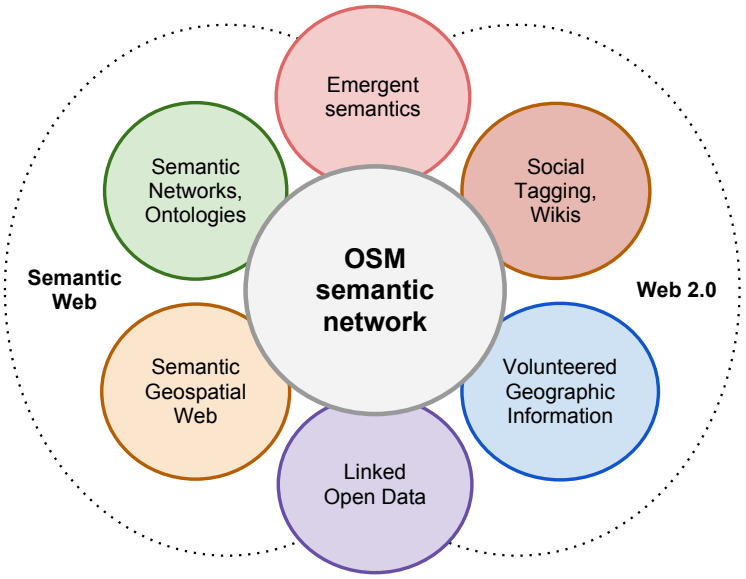

Figure 3.5: OSM Semantic Network, between Web 2.0 and the Semantic Web

order to provide support for general geospatial tasks such as computing the semantic similarity, a heuristic-based mapping with DBpedia was insufficient. Therefore, we proceeded to extract the semantic content hosted in the OSM Wiki website. This process, which resulted in the OSM Semantic Network, is outlined in the next section.

3.3 Our approach: the OSM Semantic Network

In VGI projects, contributors have to negotiate a shared set of geographic concepts. The community animating OpenStreetMap (OSM) performs most of this process on a dedicated website. To explore OSM semantics, we have developed a semantic network, representing concepts and their mutual relationships. This section outlines the salient characteristics of the OSM Semantic Network. This resource offers a machine-readable structure, describing geographic concepts commonly used in the OSM vector dataset. The OSM Semantic Network is conceived as a linkage between the Semantic Web and Web 2.0 geo-technologies (see Figure 3.5). The project is freely available online.[8]

In the OSM vector dataset, map objects are encoded as 'nodes' (points of interest or centroids), 'ways' (lines and polygons), and 'relations' (groups of objects). The world dataset currently contains 1.2 billion nodes, 106 million ways, and 1 million relations.[9] Every map object is described through properties called 'tags', defining the semantic content of the object (e.g. *amenity=university*). Such tags are proposed, defined, discussed, and sometimes discarded on the OSM Wiki website, which hosts detailed definitions

[8] `http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork`
[9] `http://wiki.openstreetmap.org/wiki/Statistics`

and usage guidelines.[10] This website is used as a reference to document and facilitate the mapping process, which is conducted through separate, dedicated web services and tools, which are outside the scope of this thesis. According to the OSM Wiki website, tagging should deliberately be informal, loose, and open. As put in the guidelines, "the only principle that really applies is KISS (Keep It Simple, Silly)."[11]

Mappers are encouraged to use well-known tags, but they are not discouraged from creating new tags when it is deemed useful. This is a more radical policy than that of comparable projects, such as Wikimapia.[12] New tags can be seen as mutations of existing tags, often discarded by the community. In this sense, tags seem to follow an evolutionary path, leading to a slow adaptation of the semantic network to the geographic reality that the vector map is supposed to represent.

The OSM keys can represent groups of geographic entities (e.g. *waterway*, *landuse*, *natural*), or encode properties with unrestricted values (e.g. *name*, *addr:street*). Although in principle any object can be included in OSM, the contributors tend to document in the OSM Wiki website only generic concepts, and not specific instances. For this reason, proper nouns are generally out of the scope of the project. Moreover, OSM favours permanent entities, and does not include entities with a short temporal lifespan such as historical events.

While some keys have a small set of well defined values (e.g. *junction*), other keys have become very large, overstretching their semantic boundaries. The key *amenity*, for example, is associated with more than 150 values, ranging from fast food restaurants to hospitals and cinemas. Moreover, similar tags can be defined in different keys, resulting in semantic difficulties for the users (e.g. *landuse=garages* versus *amenity=parking*). This semantic gap can cause disagreements among users, occasionally resulting in 'tag wars' [211].

To date, the OSM community has about 453,000 contributors. Through the OSM Wiki website, this large group negotiates what Kuhn [161] calls the 'social agreements' needed to define common semantic symbols that can be understood by most users. The fluid openness of OSM semantics is both a strength and a weakness of the project. While contributors are attracted to the lack of formal validation procedures to make changes to the map, this degree of freedom generates noise in the form of semantic ambiguity and redundancy. For this reason, several efforts have been undertaken to monitor the tag usage in the vector dataset, such as the web services TagInfo[13] and TagWatch.[14]

The OSM Wiki website encodes semantic content as a collection of interlinked pages, discussing aspects of the OSM vector dataset. Textual descriptions, images, and links to Wikipedia are used by contributors to clarify the meaning and usage of OSM tags. An OSM tag (or a key) corresponds to a

[10] `http://wiki.openstreetmap.org/wiki/Map_Features`
[11] `http://wiki.openstreetmap.org/wiki/Any_tags_you_like`
[12] `http://wikimapia.org`
[13] `http://taginfo.openstreetmap.org`
[14] `http://wiki.openstreetmap.org/wiki/Tagwatch`

shared concept, grounded in the OSM Wiki website.[15] Hence, the key idea behind the OSM Semantic Network is that the OSM Wiki can be seen as a semantic network, in which the pages are concepts and the internal links encode semantic relationships [239]. In such a network, concepts have connections with other concepts. As pages are modified and reconnected to other pages, the network topology changes accordingly. More formally, the OSM Wiki can be conceptualised as a directed graph $\mathbf{G} = (V, E)$, where vertices V are the web pages, and edges E are their hyperlinks. In order to extract a representation of this semantic network, we focused on key and tag pages in English. The OSM Wiki pages can be categorised as follows:[16]

1. **Key page.** Describes the meaning and usage of an OSM key, grouping several tags with the same key. For example the page `osmwiki:Key:amenity` summarises the key *amenity* and its recommended values (e.g. *university*, *pub*). When a key page is deemed to be too long, it generally gets split into several tag pages. To be used in the vector dataset, a key has to be associated with a value.

2. **Tag page.** Describes a specific key/value pair, representing a concept in the semantic network. For example, `osmwiki:Tag:amenity=library` defines the tag *amenity=library*. When the value is '*', the key accepts an open value (a string, a number, or a range).

3. **Proposed tag page.** Some tags have been proposed by contributors and are undergoing review. For instance, the tag *historic=aqueduct* has been proposed in `osmwiki:Proposed_features/aqueduct` and is currently marked as a draft.

4. **Cluster pages.** Pages that group related links to tag pages, while not representing directly a tag (e.g. `osmwiki:Building_attributes`).

5. **Other pages.** All the pages that do not fall in the previous categories, including contributor profiles, technical pages unrelated to tags, and administrative pages (e.g. `osmwiki:Linear_maps`).

The open source tool that we have developed, the OSM Wiki Crawler, extracts a semantic network from the OSM Wiki website, in the form of an RDF graph (see Section 4.2). The graph vertices represent OSM keys, tags, clusters, Wikipedia pages and LinkedGeoData terms. The Wikipedia and LinkedGeoData vertices have only incoming links from OSM vertices, to which they are semantically equivalent (e.g. `osmwiki:Tag:amenity=embassy` is linked to `http://en.wikipedia.org/wiki/Embassy` and to `lgdo:Embassy`). The edge labels specify a number of different relationships between vertices, ranging from links to a tag key (`osmwiki:key`) to a logical implication

[15] For this reason, in the context of OSM semantics, we use the terms 'concept' and 'tag' interchangeably.

[16] `osmwiki:` stands for the namespace `http://wiki.openstreetmap.org/wiki/`

URI	Description	Instances
Vertices		
`osmwiki:Key:`⟨*key*⟩	OSM Key.	1,503
`osmwiki:Tag:`⟨*key* = *value*⟩	OSM Tag.	2,047
`osmwiki:Proposed_features/`⟨*tag*⟩	OSM Proposed Tag.	784
`osmwiki:`⟨*page*⟩	OSM Cluster page.	22
others	LGD and Wikipedia nodes.*	2,111
Edges		
`osmwiki:link`	Internal hyperlink in OSM Wiki.	12,974
`osmwiki:key`	Link to OSM key page.	5,408
`rdf:rdf-schema#comment`	OSM Tag description.	2,889
`osmwiki:combinedWith`	Tag is combined with target tag.	2,054
`osmwiki:wikipediaLink`	A link to a Wikipedia page.	1,604
`owl:owl#equivalentClass`	Equivalent class in other ontology.	652
`osmwiki:implies`	Tag implies target tag.	226

Table 3.1: OSM Semantic Network vertices (total: 6,467) and edges, sorted by number of instances (total: 28,807); (*) leaf vertices. Graph extracted on February 1, 2012.

(`osmwiki:implies`). An important part of the network is that of lexical definitions used by users to describe tags. A sample of such definitions is displayed in Table 3.2.

Generic internal hyperlinks (`osmwiki:link`) are particularly important, as they capture general relatedness between the source and the target pages, useful to compute a cognitively plausible semantic similarity. For example, *amenity=library* contains generic links to *tourism=museum* and *shop=books*. Besides, cluster pages do not represent tags directly, but contribute to the modelling of the semantic similarity between tags. For instance, the cluster page `Building_attributes` strengthens the connectivity between several tags related to buildings.[17] The detailed content of the RDF graph is presented in Table 3.1. An example of the network structure is displayed Figure 3.6, which represents the relationships between the concepts *university*, *school*, and *building*. The implementation of the OSM Wiki Crawler is described in detail in Section 4.2. Pre-extracted networks are available online.[18]

The OSM Semantic Network contains valuable, machine-readable semantic information about geographic concepts utilised in OSM. A key area of application for this semantic resource is the computation of semantic relatedness and similarity, supporting a number of applications, including GIR, map personalisation, and information integration. To clarify the nature and scope of our contribution, in the next section we define the difference between two conceptions of geographic data semantics.

3.4 Weak and strong geo-semantics

The OSM Semantic Network aims at filling a gap between Web 2.0 and the Semantic Web, also known as 'Web 3.0.' In order to clarify our contribution to

[17] `osmwiki:Proposed_features/Building_attributes`
[18] `http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork`

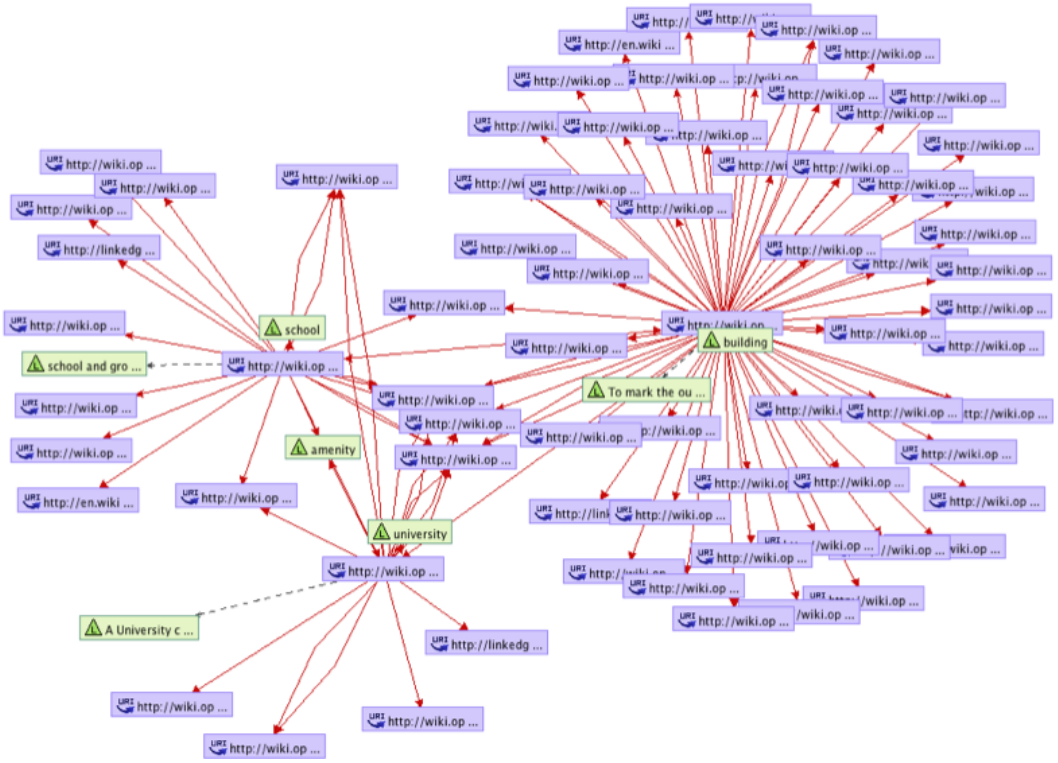

Figure 3.6: OSM Semantic Network: subset representing concepts *university*, *school*, and *building*.

VGI and OSM, this section proposes a distinction between 'strong' and 'weak' approaches to the complex issue of geo-semantics. This is particularly important in clarifying our aims in relation to the broad vision of the Semantic Web [31], and its geographic branch, the Semantic Geospatial Web [65].

In the heated debate surrounding the Semantic Web, it is beneficial to go beyond the dichotomy between the hype fuelled by its supporters, and the scepticism of its radical critics. The original vision of the Semantic Web implies a form of automatic *understanding* of data, which echoes the early artificial intelligence (AI) programme, i.e. implementing machine intelligence through predicate-based languages and inferential engines [121, 253]. On this ground, critics attacked the Semantic Web's core tenets, describing them as utopian and, at best, over-optimistic [62, 82]. Obviously, a similar criticism may be levelled at the Semantic Geospatial Web in the geographic domain. While it is reasonable to envisage a *syntactic* standardisation of geographic data promoted for example by the Open Geospatial Consortium (OGC), an effort to homogenise the *semantics* of data and services is certainly more challenging (see Section 2.4.3 for a review on current trends in geo-semantics).

American philosopher John Searle [264] outlined the distinction between 'weak AI' and 'strong AI.' Weak AI, he argued, sees digital computers as powerful tools that can mimic certain human cognitive tasks. Strong AI, by contrast, assumes that computers are equivalent to the human mind, claiming the theoretical possibility of engineering human-like *understanding* in a computer programme – a possibility that Searle firmly rejects. More recently, Obrst

[224] has proposed a spectrum to classify semantic technologies, ranging from 'weak semantics,' such as plain ER schema, to 'strong semantics,' e.g. full-fledged ontologies with inferential power. A similar distinction was proposed in the domain of Web-based adaptive systems by Torre [289]. She defined a 'weak semantic web,' consisting of "adding meaning to resources to improve services to users," contrasted with a 'strong semantic web,' which aims at "data understanding by machines to improve services to users" (p. 437). Hence, taking inspiration from these proposals, we define a 'geo-semantic spectrum,' at whose extremes lie a 'strong' and a 'weak' approach to geo-semantics:

Weak geo-semantics. The meaning of spatial data is described using semantic markers, i.e. unstructured raw text, labels, and tags. For example, the tag 'lake' can be associated with a bi-dimensional polygon. In weak geo-semantics, the meaning of map objects is left *implicit*, delegating all the cognition to human consumers. Weak geo-semantics techniques exploit such simple meta-data to retrieve, manipulate, and utilise geographic information. Simple key word-based search mechanisms can be seen as a basic form of weak semantics: the map object is described with a symbol (e.g. 'lake Garda'), and the user relies on that symbol to retrieve it. Overall, weak geo-semantics is simple, highly flexible, robust, scalable and dynamic, but is limited by the lack of formal structure and by the ambiguity of natural language. Weak semantic data is difficult to integrate with other data sources, and to re-use in different contexts. Overall, the 'intelligence' of the system tends to be low. VGI and OSM rely mostly on weak geo-semantic models, such as collaborative tagging and folksonomies [297].

Strong geo-semantics. The semantics of geographic data is represented *explicitly*, using full-fledged formal ontologies, which capture salient aspects of the domain in a coherent way [110]. Strong geo-semantics rely on predicate-based languages from logic and AI, such as Resource Description Framework (RDF), Web Ontology Language (OWL), description logic (DL), and first-order logic. New knowledge can then be generated using inferential mechanisms, and diverse services and datasets can inter-operate. Following the Semantic Web tenets, strong semantics posits the possibility of grounding the meaning of geographic data on such formalisms, providing a 'shared semantic ground' for data integration [161]. Strong geo-semantic technologies are complex, highly structured, fragile, and have a certain inferential power – within the limits of the specific formalisms being employed. This fact makes their adoption impractical in noisy and dynamic environments such as VGI. Strong semantics can support advanced geospatial tasks, such as information integration from heterogeneous sources, map personalisation, implicit feedback, and GIR, increasing the 'geographic intelligence' of the systems. In its extreme form, strong geo-semantics is equivalent to strong AI, and may therefore be regarded with scepticism [82].

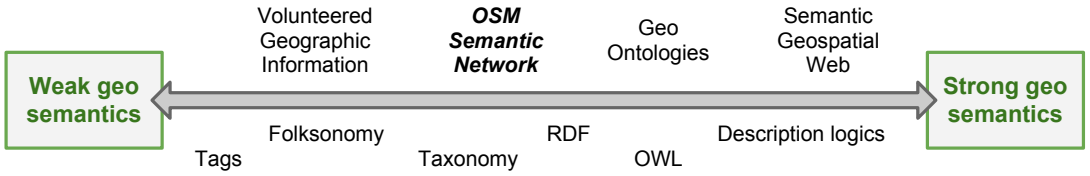

Figure 3.7: The geo-semantic spectrum

Figure 3.7 depicts the geo-semantic spectrum, locating the OSM Semantic Network between the extremes of strong and weak geo-semantics. It can be argued that most real-world applications in the GeoWeb lie somewhere on the spectrum: VGI tends to fall closer to the 'weak' area, while sophisticated technologies for the Semantic Geospatial Web aim at reaching the 'strong' extreme. The geo-semantic spectrum can be interpreted as a trade-off between the model's implicit and explicit semantics: more explicit semantics results in higher complexity, and vice-versa. Our contribution to OSM aims at enabling 'stronger' semantic support for OSM, admitting the difficulties of building a strong semantic model for its vector dataset. Therefore, the OSM Semantic Network can be seen as pushing OSM towards the strong end of the semantic spectrum, while still remaining in the intermediate area.

Regardless of one's position towards the Semantic Web as a global endeavour, it is undeniable that advancing semantic support for geo-data can greatly benefit the search, manipulation, and usage of VGI in practical circumstances. Several industrial applications expressing data in RDF indicate that, beyond the hype, 'stronger' semantics can actually support real applications [42]. In this context, we think that measures of semantic relatedness and similarity can benefit a range of semantic technologies for OSM, helping move the applications towards a stronger semantics. The next section provides our definitions of semantic relatedness and similarity.

3.5 Semantic relatedness and similarity

This section aims at defining two aspects of similarity and semantics that are at the centre of this thesis, semantic relatedness and semantic similarity, considered in the context of geographic information. The existing definitions for semantic relatedness and similarity were discussed thoroughly in Section 2.5.1. In our work on OSM data, semantics has arisen as a key area to provide support for many geographic applications.

Semantic relatedness and similarity are paramount aspects of the complex interplay between geographic information and human experience. As discussed throughout Section 2.5, computing the semantic similarity of geographic objects can help develop successful approaches to a vast range of problems, including GIR, data mining, information integration, and spatial recommender systems. For example, if a user often examines ports and sea routes, a system should be able to infer a general interest in naval transport, and model

Key/Value	Lexical definition from OSM Wiki website
tourism=information	An information source for tourists, travellers and visitors. May include: Tourist information centres and offices. Map boards, such as town maps, site maps, hiking and other sport trail maps. ...
sport=athletics	Track and field athletics. A collection of sports events that involve running, throwing and jumping.
amenity=taxi	A place where taxis wait for passengers. Often found at airports, hotels, railway/bus/subway stations, large shopping centers. Approved.
leisure=garden	Place where flowers and other plants are grown in a decorative and structured manner or for scientific purposes. ... A garden can have aesthetic, functional, and recreational uses: Not to be confused with *shop=garden_centre*.
building=university	A university building. Completed with the tag *amenity=university*.
*room=**	room marks a room inside a building. Man made. The room Key is used to mark nodes or areas as a room, typically inside a building. Part of Proposed features/indoor.
natural=wetland	An area subject to inundation by water, or where waterlogged ground may be present. Examples include swamps, tropical mangroves, tidal, marshland, and bogs. Wetlands are also typically found at the fringes of rivers, lakes or along the coastline.

Table 3.2: Examples of lexical definitions of OSM tags

that piece of information in her user profile, to extend recommendations beyond a narrow boolean model. For this reason, we devoted efforts to devise NetLexSiM, an approach to computing the semantic similarity in the context of OSM volunteered geographic concepts, tapping the knowledge encoded in the OSM Semantic Network. In the next two sections we propose a definition of semantic relatedness and similarity, in the framework of geo-semantics.

3.5.1 Semantic relatedness

Two geographic concepts are semantically related to the degree to which they co-occur in human experience. In the OSM Semantic Network, related concepts co-occur in concept definitions. The spatial aspect of relatedness is captured by Tobler's first law of geography, which asserts that everything is related to everything else, but near things are more related than distant things [288]. In the context of geo-semantics, we recast Tobler's law as follows:

> *Every concept is semantically related to all other concepts, but concepts used to describe the same domain of human experience are more related than other concepts.*

This law puts geographic concepts in relation to human spatial experience from which concepts arise, suggesting indistinct, gradual, and shifting boundaries between related and unrelated concepts. Semantic relatedness is intrinsically fuzzy, admitting a continuous spectrum of relatedness rather than

a binary classification (i.e. *related* or *unrelated*). Highly related concepts belong to the same semantic field. The same concepts can belong to n overlapping semantic fields. Relatedness includes all semantic relations, including synonymy, antonymy, hyponymy, hypernymy, holonymy, meronymy, causality, temporal contiguity, function, proximity, and containment.

3.5.2 Semantic similarity

While semantic relatedness of concepts can be defined based on co-occurrence in human experience, semantic similarity of geographic concepts can be only determined through the analysis of the concepts' characteristics. The complexities of psychological and cognitive research on similarity were reviewed in Section 2.5. In the OSM Semantic Network, concepts are described through lexical definitions, and links to other concepts, and techniques to compute similarity can exploit these aspects.

In general, semantic similarity is a subset of semantic relatedness, i.e. all similar concepts are also related, but related concepts are not necessarily similar. From a linguistic perspective, semantic similarity includes synonymy, hyponymy, and hypernymy. In the context of VGI, we have extracted the OSM Semantic Network, which consists of an explicit, crowdsourced categorisation of geographic knowledge. In order to share concepts, contributors describe them with lexical definitions, and by inter-linking them. For example, to describe the concepts 'stream', 'canal' and 'river', users employ concepts such as 'water', 'waterway', etc. In a circular reference, 'waterway' is described by using concepts 'stream', 'river', and 'canal'.

Adopting a bag-of-words (BOW) model, a concept definition contains semantically related concepts, and, recursively, similar concepts are described using similar concepts. By combining Tobler's law with Jeh and Widom's recursive definition of topological similarity [136], we define semantic similarity in the following way:

> *All concepts are semantically similar, but concepts defined with similar concepts are more similar than other concepts.*

The work described in this thesis focuses on techniques for semantic similarity. Through the survey described in Section 6.2, we collected the Geo Relatedness and Similarity Dataset (GeReSiD), a dataset containing human judgements on the semantic relatedness and similarity of geographic concepts. In this way, it is possible to determine empirically to what degree a technique approximates semantic similarity and relatedness.

Having provided a definition of semantic relatedness and similarity tailored to the geographic domain, the next sections outline our approach to semantic similarity, NetLexSiM, which combines a co-citation based approach to similarity (Section 3.6), and an approach to lexical similarity (Section 3.7).

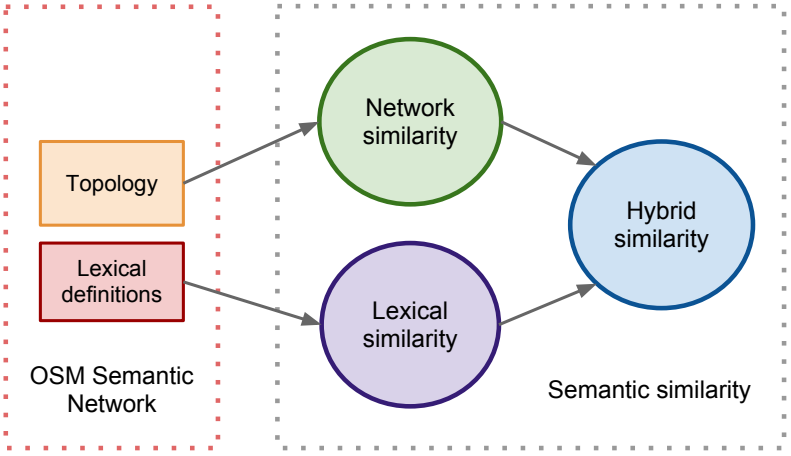

Figure 3.8: The structure of NetLexSiM

3.6 NetLexSiM: network similarity

This section outlines the first component of NetLexSiM, our approach to computing a semantic similarity in the OSM Semantic Network, i.e. the network-based, topological similarity. The second component, the lexical similarity, exploits the lexical definition of concepts, and will be described in Section 3.7. The two components, network and lexical, are combined into a hybrid measure. The general structure of NetLexSiM is depicted in Figure 3.8.

In the OSM Semantic Network, vertices represent concepts, and edges their mutual connections. To exploit the semantic content of the network, we explore its potential to compute the semantic similarity of OSM concepts. We define a similarity measure between the tags a and b as $s(a, b) \in [0, 1]$, where 0 means no similarity, and 1 means maximum similarity being a and b vertices in the OSM Semantic Network. Network-based similarity techniques assume that the relationships between concepts must be sufficiently rich and representative [261]. To assess whether its dense link structure contains valid knowledge about the OSM tags, we compute a similarity score based purely on the network topology, ignoring the lexical descriptions of concepts.

Approaches such as MDSM and Sim-DL have been devised specifically for geographic concepts (see Section 2.5.2). However, because such measures require the data to be expressed in specific formalisms – a detailed description of attributes, parts, and roles, or DL – they cannot be used in the context of OSM concepts. Relevant similarity measures are those devised on the semantic network WordNet, reviewed in detail in Section 2.5.4. The OSM Semantic Network does not contain a deep taxonomy such as that of nouns and verbs in WordNet, but is a densely connected graph. The WordNet-based measures are tailored to its topology, and our semantic network is too different to allow a direct application of such measures on it.

The OSM Semantic Network reflects the shallow key/value structure of OSM semantics. Additionally, because of the shallow structure of OSM se-

Symbol	Description
$\mathbf{G} = (V, E)$	the directed graph in which each vertex $a \in V$ represents a OSM tag and $\langle a, b \rangle \in E$ is a hyperlink from tag a to b.
$s(a, b)$	similarity score between tags a and $b \in V$. $s(a, b) \in [0, 1]$, $s(a, b) = s(b, a)$. When $a = b$, $s(a, b) = 1$.
$I(a)$	set of incoming links to tag $a \in V$. $\|I(a)\|$ is the indegree of a.
$O(a)$	set of outgoing links to tag $a \in V$. $\|O(a)\|$ is the outdegree of a.
C	P-Rank decay factor. $C \in (0, 1)$. If $C = 1$, P-Rank does not converge.
λ	P-Rank in-out balance constant. $\lambda \in [0, 1]$. $\lambda = 1$: incoming links; $\lambda = 0$: outgoing links.
k	P-Rank current iteration. $k \in [0, K]$.
K	P-Rank maximum iterations. $K \in [1, \infty)$.
$\mathbf{R}_k$	P-Rank score matrix at iteration k.
$\mathbf{T}_i$	transition matrix of $\mathbf{G}$ constructed on $I(a)$.
$\mathbf{T}_o$	transition matrix of $\mathbf{G}$ constructed on $O(a)$.
$\mathbf{\Theta}$	diagonal matrix. $\forall k$, when $a = b$, $\mathbf{\Theta}(a, b) + \mathbf{R}_k(a, b) = 1$.

Table 3.3: Notations for network-based similarity

mantics, the paths between OSM tags are very short: the majority of concepts are connected through 2 edge-paths, even when semantically very dissimilar (e.g. *sauna* → *amenity* → *bench*). Shortest-path based techniques need paths of variable lengths to be effective, and are therefore doomed to fail in this case. This aspect is visible in the LinkedGeoData and OSMonto ontologies too (see 2.3.3). To compute the semantic similarity of OSM tags it is necessary to identify alternative measures. The co-citation approach seems promising.

As discussed in Section 2.5.3, co-citation algorithms aim at finding similarity in a graph of inter-linked objects, based on the intuition that similar objects are referenced together. Although it is possible to compute co-citation measures on the LinkedGeoData and OSMonto ontologies, this would result in a binary classification between tags that are in the same subtree (e.g. *amenity=school*, *amenity=fountain*) or not (e.g. *amenity=school*, *landuse=forest*). This approach is unable to account for semantic similarity within the same key, e.g. *amenity=school* and *amenity=university* are expected to be more similar than *amenity=school* and *amenity=fountain*. On the other hand, our OSM Semantic Network allows for a finer computation of similarity by including general hyperlinks between pages, and can distinguish between these cases.

3.6.1 P-Rank

To the best of our knowledge, co-citation algorithms have not been utilised to compute semantic similarity of geographic classes. To fill this knowledge gap, we decided to investigate whether they would be suitable to compute the similarity of concepts in the OSM Semantic Network. Hence, we considered

P-Rank, a highly generic co-citation algorithm devised by Zhao et al. [318]. By setting different values to its parameters, P-Rank is equivalent to earlier algorithms, including Co-citation [271], Coupling [152], and Amsler [9], SimRank [136], and rvs-SimRank [318]. For this reason, it is possible to observe the performance of co-citation algorithms by exploring the result space of P-Rank. In this context, we propose a linear algebra formulation of P-Rank, discussing in detail the meaning and impact of its parameters (K, λ, and C), largely left implicit in the literature [318, 136, 178].

P-Rank is a recursive measure of similarity, based on the combination of two recursive assumptions: (1) two entities are similar if they are referenced by similar entities; (2) two entities are similar if they reference similar entities. Our recursive definition of semantic similarity, i.e. similar concepts are described using similar concepts, is based on these assumptions (see Section 3.5). All of the notations and symbols used in this Section are summarised in Table 3.3. $\mathbf{G} = (V, E)$ is a directed graph, where each vertex $v \in V$ represents a concept and $\langle u, v\rangle \in E$ is a semantic relation from concept u to v. For each concept v, function $I(v)$ denotes incoming links ($\langle u, v\rangle \in E$), while $O(v)$ the outgoing links ($\langle v, u\rangle \in E$). $|I(v)|$ and $|O(v)|$ denote respectively the number of incoming and outgoing links to concept v. In the case of the similarity of the same object, the P-Rank score is 1 ($s(a, b) = 1$ when $a = b$). In all other cases, when $a \neq b$, the P-Rank similarity score is shown in the recursive Formula 3.4.

$$s(a,b) = \lambda \times \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b)) + (1-\lambda) \times \frac{C}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} s(O_i(a), O_j(b)) \tag{3.4}$$

When $|I(a)||I(b)| = 0$ or $|O(a)||O(b)| = 0$, the respective expression is nullified. If $|I(a)||I(b)| = 0$ and $|O(a)||O(b)| = 0$, then $s(a, b) = 0$. Conceptually, the similarity can be seen as flowing in a graph of vertex pairs $\mathbf{G^2} = (V^2, E^2)$, in which an edge $\langle (a, b), (c, d)\rangle$ exists if and only if edges $\langle a, c\rangle$ and $\langle b, d\rangle$ exist in E.

P-Rank is calculated iteratively, choosing a number of iterations $K \in [1, \infty)$. The higher K, the better the approximation of the theoretical solution to P-Rank. At the first iteration R_0 ($k = 0$), the scores are initialised to 0, $R_0(a, b) = 0$, apart from the identities (if $a = b$, then $R_0(a, b) = 1$). All P-Rank iterations with $k > 0$ can be expressed as a series of iterations converging to the theoretical similarity score:

$$s(a,b) = \lim_{k\to\infty} \mathbf{R}_k(a,b) \tag{3.5}$$

$$\mathbf{R}_k = C(\lambda \cdot \mathbf{T}_i \mathbf{R}_{k-1} \mathbf{T}_i' + (1-\lambda) \cdot \mathbf{T}_o \mathbf{R}_{k-1} \mathbf{T}_o') + \mathbf{\Theta} \tag{3.6}$$

The similarity $s(a, b)$ is a function $f(C, \lambda)$. The constant C is the decay factor applied to the recursive propagation of similarity across the edges. When

C is close to 0, almost no similarity flows from one pair to its neighbours, while with C close to 1 the opposite situation arises. The constant λ, on the other hand, is the in-outlinks balance. When $\lambda = 1$, only incoming links are considered, while $\lambda = 0$ indicates that the similarity is computed only on the outgoing links.

The number of iterations K determines the minimum precision of the algorithm, i.e. the maximum gap between $s(a, b)$ and $\mathbf{R}_k(a, b)$, which decreases as K grows [183]. K, while obviously influencing $\mathbf{R}_k(a, b)$, has no impact on $s(a, b)$. In order to establish the accuracy of P-Rank to the k-th iteration, Lizorkin et al. [183] assessed that the difference between the theoretical similarity ($k = \infty$) and its k-th iteration is always smaller or equal than C^{k+1}:

$$s(a, b) - \mathbf{R}_k(a, b) \leq C^{k+1} \tag{3.7}$$

By applying P-Rank on the OSM Semantic Network, it is possible to observe the behaviour of a number of co-citation algorithms. The results of this evaluation are reported in Sections 5.5 and 6.3.

3.6.2 Complexity of network similarity

This section discusses the computational complexity of the co-citation-based approach to network semantic similarity in the OSM Semantic Network. Non-recursive co-citation algorithms visit all the incoming and/or outgoing links for a pair of vertices (a, b) [271, 152, 9]. These approaches result in the linear complexity $O(deg(a) + deg(b)) \leq O(n)$, where deg is the degree of a vertex, i.e. the number of incident edges. However, non-recursive co-citation tends to reach modest results, and is largely outperformed by recursive approaches, such as SimRank, and P-Rank [136, 318]. For a graph containing n vertices, the spatial complexity of SimRank is $O(n^2)$, i.e. a similarity matrix to be updated at each iteration. The temporal complexity of the algorithm is remarkably higher, being $O(K \cdot n^2 \cdot mean(deg)^2) \leq O(K \cdot n^4)$, where K is the number of iteration, and $mean(deg)$ the mean degree of the graph.

Given the promising results but high spatio-temporal complexity of recursive co-citation techniques, several efforts have been undertaken to optimise them. Lizorkin et al. [183] reduced the SimRank complexity to $O(K \cdot n^3)$, while Yu et al. [317] devised an approximation of SimRank whose temporal complexity is further reduced to $O(n^3 + K \cdot n^2)$. Li et al. [178] offered an approach to computing the similarity of a vertex pair, without computing the whole similarity matrix of size n^2.

These improvements notwithstanding, recursive co-citation techniques remain very complex. However, given the limited size of the OSM Semantic Network (about 6,500 vertices), the complete similarity matrix of size $\approx 4 \cdot 10^7$ can be easily computed, stored, and indexed on regular hardware. Such precomputed similarity matrices can then be used in a number of applications, for example to retrieve features conceptually similar to a given feature, to recommend semantically similar map features in a spatial recommender system, or to personalise a map based on the user's semantic interests.

3.7 NetLexSiM: lexical similarity

This section outlines the second component of NetLexSiM, our approach to computing semantic similarity in the OSM Semantic Network: the lexical similarity between concept definitions, expressed in natural language by contributors. The OSM Semantic Network represents geographic concepts as vertices in a graph (see Section 3.3). Most concepts are also described through a lexical definition, extracted from the OSM Wiki website.

OSM contributors define, update, and refer to these definitions to create and utilise the OSM vector dataset [145]. Table 3.2 shows examples of these lexical definitions. Depending on the concept, definitions can be very detailed and redundant (e.g. *leisure=garden*), or extremely concise, when no further explanation is deemed necessary (e.g. *building=university*). Some definitions contain OSM-specific details (e.g. 'Part of Proposed features'), while others are similar to encyclopedic definitions of general map entities (a sample of definitions was reported in Table 3.2, in Section 3.3). In this section, we describe a knowledge-based approach to computing the semantic similarity of such lexical definitions, relying heavily on the semantic network WordNet [74].

It is possible to use tags themselves as very short lexical descriptions of concepts. However, this approach results in high ambiguity, beyond the intrinsic vagueness of geographic concepts[298]. For instance, *tourism=information* indicates an information resource for tourists. The term 'tourism' is very specific and semantically unambiguous, i.e. it has only one sense of WordNet, "the business of providing services to tourists" (*tourism#n#1*).[19] The noun 'information,' on the other hand, has 5 senses, including "a message received and understood" (*information#n#1*), "knowledge acquired through study or experience or instruction" (*information#n#2*), and "formal accusation of a crime" (*information#n#3*).

Therefore, a simple direct match between OSM tags and WordNet synsets (or similar lexical ontologies) results in high ambiguity, and obtains low cognitive plausibility. Unsurprisingly, the full lexical definition contributes to substantially decrease the ambiguity of the tag. The definition of *tourism=information*, for example, offers a rich set of terms that are useful to narrow down the concepts related to that map entity: among others, it refers to 'tourists', 'travellers', 'hiking', 'history', 'wildlife', etc.

In this context, the goal is to measure the semantic similarity between two OSM concepts uniquely on these lexical definitions, ignoring any other information available in the OSM Semantic Network. To achieve this, we propose an unsupervised method based on the extraction of semantic terms, then combined in vectors (see Table 3.4 for notations):

Semantic terms. Semantic terms are obtained from the concept lexical definitions, expressed in natural language by OSM users. Semantic terms are tagged with part-of-speech (POS) tags, in order to identify nouns and

[19] All the definitions and word senses have been taken from WordNet 3.1.

Symbol	Description.
t_a	Semantic term used to describe concept a. It can be a noun (e.g. *lake*, *wetland*), a verb (e.g. *eat*), or an adjective (e.g. *large*). Each t is also described by a part-of-speech tag (*JJ* for adjectives, *NN* for nouns, *VB* for verbs). $t_a = \langle word, POS \rangle$
C	A corpus of text documents. The corpus is utilised to extract weights of semantic terms w_{tC}.
w_{tC}	Semantic weight of term t in corpus C. $w \in [0, 1]$
a	A geographic concept, described by a textual definition in natural language (e.g. *landuse=wetland*: "An area subject to inundation by water, or where waterlogged ground may be present. ...").
$\vec{a}$	Semantic vector representing concept a. $\vec{a} = w_{a1} \cdot t_{a1} + \ldots + w_{an} \cdot t_{an}$ $\sum w = 1$
$sim_t(t_a, t_b)$	Semantic similarity measure between terms t_a and t_b. $sim_t(t_a, t_b) \in [0, 1]$. $sim_t(t_a, t_b) = sim_t(t_b, t_a)$
$\hat{s}(t_i, \vec{v})$	Maximum similarity of term t_i against terms in vector $\vec{v}$.
$sim'_v(\vec{a}, \vec{b})$	Asymmetric similarity measure between semantic vectors $\vec{a}$ and $\vec{b}$. $sim'_v(a, b) \in [0, 1]$, $sim'_v(a, b) \neq sim'_v(b, a)$.
$sim_v(\vec{a}, \vec{b})$	Symmetric similarity measure between concepts a and b, derived from $sim'_v(\vec{a}, \vec{b})$ and $sim'_v(\vec{b}, \vec{a})$. $sim_v(\vec{a}, \vec{b}) \in [0, 1]$, $sim_v(\vec{a}, \vec{b}) = sim_v(\vec{b}, \vec{a})$. $sim_v \leftarrow f(sim_t)$

Table 3.4: Notations for lexical similarity

verbs, discarding all the other parts of speech. Given a lexical corpus C, weights w_{tC} for each term are extracted – Inverse Document Frequency (IDF) is a common weighting scheme [140, 254]. Term-to-term semantic similarities are then computed through function sim_t (Section 3.7.1).

Semantic vectors. A semantic vector $\vec{a}$ for a geographic concept a is constructed from the content words. The vector terms are weighted through the corpus weights w_{tC}. The similarity of concept vectors $\vec{a}$ and $\vec{b}$ is computed through a vector-to-vector function sim_v, based on the techniques described in Corley and Mihalcea [54] and Fernando and Stevenson [75] (Section 3.7.2).

This method is based on a Vector Space Model, and is informed by the statistical semantics hypothesis, which states that statistical patterns in word usage can reveal information on word meaning [294]. The remainder of this section describes in detail the steps to compute the semantic similarity of two geographic concepts a and b.

3.7.1 Semantic terms

The definition of a geographic concept a consists of a portion of text in natural language. The building block of a lexical definition is a semantic term

t, intended as a content word contributing to the overall meaning of the concept. In linguistics, 'content words' are "lexical items which have a relatively 'stable and detailed' semantic content and as such carry the principal meaning of a sentence" [55, p. 1]. 'Function words,' on the other hand, fulfill a grammatical role, describing roles and relationships between content words. For example, given the definition "place where flowers and other plants are grown", the content words 'place,' 'flowers,' and 'plants' are more important than the functional words 'where,' 'and,' and 'other' to understand that the concept being defined is a garden.

A semantic term t is also described by a part-of-speech tag ($t = word/POS$), following the format of the Penn Treebank [195]. Among all the Penn tags, we restrict our method to nouns (NN) and verbs (VB). Adjectives (JJ) were originally included, but computing their similarity proved problematic, and they were finally excluded. To facilitate their usage, the terms have to be lemmatised (e.g. 'rivers' to 'river,' 'mice' to 'mouse,' 'ate' to 'eat,' etc.). Hence, the aforementioned definition contains the semantic terms $place/NN$, $flower/NN$, $plant/NN$, and $grown/VB$.

In the definition of concept a, semantic terms can contribute to its meaning in different ways. When defining the tag *building=university*, the terms *university/NN* and *building/NN* carry more meaning than *completed/JJ* or *tag/NN*. Thus, an appropriate weighting scheme is used to compute the 'semantic weight' of a term t in the corpus (w_{tC}). The semantic weight represents how important a term is in conveying the meaning of a definition. A relatively infrequent term in a corpus is expected to be lighter than a very frequent one.

A corpus of text documents C is utilised to compute the semantic weights of terms on a statistical basis. The corpus C is a set of text documents, each containing semantic terms. Formally, $d_i = \{t_0, t_1 \dots t_n\}$ and $C = \{d_0, d_1 \dots d_m\}$. A term can be non-unique ($\exists(t_i, t_j) : t_i = t_j$). A seminal approach to computing the semantic weight of the terms in a corpus is the Term Frequency-Inverse Document Frequency (TF-IDF) [140, 254, 116]. IDF scores represent how common a term t is in the whole corpus, while the term frequency (TF) is the number of occurrences of t in d:

$$idf(t, C) = \log \frac{|C|}{|d \in C : t \in d|} \qquad tf(t, d) = |\{t \in d\}| \tag{3.8}$$

Treating a definition a as a document, the semantic weight of each term t_a in a single document can easily be computed with TF-IDF:

$$tfidf(t, d, C) = tf(t, d) \cdot idf(t, C) \tag{3.9}$$

Given the wide range of applications of TF-IDF, other weighting schemes have been devised along similar lines, such as the Okapi BM25 [247].

Once the semantic terms have been collected, their semantic similarity has to be computed. In a boolean comparison, two semantic terms t_1 and t_2 can be deemed to be equal or non equal. This comparison is enough in several applications, but in the case of OSM definitions a more sophisticated ap-

proach to semantic similarity is needed. Other semantic relationships should be considered when comparing terms, such as synonyms (e.g. 'buy' and 'purchase'), hyponyms, and hypernyms (e.g. 'oak' and 'tree'). In order to compute semantic similarity exploiting relationships between terms, knowledge-based approaches are generally deemed to have higher precision than corpus-based techniques (see Section 2.5.4) [204, 1]. We define a normalised, symmetric term-to-term semantic similarity measure sim_t:

$$sim_t(t_i, t_j) \in [0, 1] \qquad sim_t(t_i, t_j) = sim_t(t_j, t_i) \qquad i = j : sim_t(t_i, t_j) = 1 \tag{3.10}$$

A wide variety of approaches can be adopted to compute the function sim_t. In particular, the lexical semantic network WordNet has been thoroughly investigated as a tool to compute term-to-term semantic similarity [74, 204]. Section 2.5.4 surveys ten WordNet similarity techniques, which are utilised in this phase.

3.7.2 Semantic vectors

A geographic concept a can be represented as a semantic vector. The semantic terms t_a contribute to conveying the meaning of a. In this sense, a is a set of semantic terms. This representation is often called bag-of-words (BOW), as it does not take into consideration the syntactical structure of text, but only the presence of terms, regardless of their order [294].

$$a = \{t_1 \dots t_n\} \qquad t_i = label, POS \tag{3.11}$$

To represent the geographic concept a in the semantic vector space, we define a multidimensional vector $\vec{a}$, whose scalars are non-negative weights, and components are semantic terms t:

$$\vec{a} = w_{a1} \cdot t_{a1} + \ldots + w_{an} \cdot t_{an} \qquad \forall w : w \in R_{\geq 0} \sum w = 1 \tag{3.12}$$

The semantic weights w convey the importance of each term in the overall description of the concept. In a vector space, the similarity of two vectors can be easily computed through Euclidean or cosine distance, and the Jaccard's index [286]. However, these approaches can succeed only when the vectors present a high overlap, i.e. shared terms. When vectors share very few terms, the similarity computation cannot capture finer nuances, conveyed for example through synonyms or conceptually similar terms. For this reason, a boolean comparison between vector terms is not satisfactory, and we utilise a range of term-to-term semantic similarity measures, described in the next section.

Given two semantic vectors $\vec{a}$ and $\vec{b}$, the semantic similarity of the corresponding concepts a and b needs to be computed. Vector similarity is traditionally computed as the inverse of the distance between the vectors, such as the Euclidean, cosine, Chebyshev, and Manhattan distances [177]. However,

these techniques are effective only when there is an overlap between the terms. When there is little or no overlap in the vectors, the vast majority of similarity scores is 0. For this reason, the term-to-term similarity function sim_t is useful to capture a fuzzy semantic distance between terms in $\vec{a}$ and $\vec{b}$.

The fuzzy comparison of two semantic vectors is in many ways analogous to paraphrase detection, i.e. recognising semantically similar sentences [54, 75]. For example, the news sentences *"The Hartford Courant reported that Tony Bryant said two friends were the killers"* and *"A lawyer for Skakel says there is a claim that the murder was carried out by two friends of one of Skakel's school classmates, Tony Bryan"* convey very similar semantic content, but show few overlapping terms [63, p. 354]. For this reason, we consider techniques originally developed in the area of paraphrase recognition/detection.

Starting from the approach presented by Corley and Mihalcea [54], the asymmetric semantic similarity between concepts a and b can be defined as the semantic similarity between their corresponding vectors $\vec{a}$ and $\vec{b}$, formalised as follows:

$$\begin{aligned} sim'_v(a,b) = sim'_v(\vec{a},\vec{b}) &= \sum_{i=1}^{|\vec{a}|} w_{ai} \cdot \hat{s}(t_{ai},\vec{b}) \\ sim'_v(b,a) = sim'_v(\vec{b},\vec{a}) &= \sum_{i=1}^{|\vec{b}|} w_{bi} \cdot \hat{s}(t_{bi},\vec{a}) \\ sim'_v(a,b) \neq sim'_v(b,a), \qquad & sim'_v(a,b) \in [0,1] \end{aligned} \tag{3.13}$$

The function $\hat{s}$ is the maximum semantic similarity score between a term t_i and a vector $\vec{v}$:

$$\begin{aligned} \hat{s}(t_i,\vec{v}) \in [0,1] \qquad \forall t_v \in \vec{v} : sim_t(t_i,t_v) \leq \hat{s}(t_i,\vec{v}) \\ sim_t(t_i,t_j) \in [0,1] \end{aligned} \tag{3.14}$$

More recently, Fernando and Stevenson [75] have developed a variant of Corley and Mihalcea's approach. Instead of considering only the terms with maximum similarity to the vector being analysed, $\hat{s}$ is the sum of all similarities.

$$\hat{s}(t_i,\vec{v}) = \sum_{j=0}^{|\vec{v}|} sim_t(t_i,t_j) \qquad t_j \in \vec{v} \tag{3.15}$$

The approach by Fernando and Stevenson [75] is expressed in linear algebra as follows, where M is a similarity matrix $i \times j$ constructed on function sim_t:

$$sim_v(\vec{a},\vec{b}) = \vec{a}M\vec{b}^T \qquad \forall i,j : M_{ij} = sim_t(t_i,t_j) \tag{3.16}$$

To compute the concept-to-concept similarity $s(a,b)$, an appropriate term-to-term similarity measure $sim_t(t_i,t_j)$ has to be utilised. As the evaluation will confirm, the choice of sim_t impacts on the cognitive plausibility of

the measure. In this sense, sim_v is a function $f(sim_t)$. While the similarity measure defined in Equation 3.13 is asymmetric, a symmetric measure sim'_v can be easily obtained as:

$$sim_v(a,b) = \frac{sim'_v(a,b) + sim'_v(b,a)}{2} \qquad (3.17)$$
$$sim_v(a,b) \in [0,1], \quad sim_v(a,b) = sim_v(b,a)$$

This knowledge-based approach based on semantic vectors enables the computation of the semantic similarity of segments of text describing concepts, contained in the OSM Semantic Network. The technique is thoroughly evaluated in Sections 5.6 and 6.4.

3.7.3 Complexity of lexical similarity

This section discusses the computational complexity of our approach to computing lexical semantic similarity for OSM concepts. Our approach is performed through the following steps: (a) extraction of semantic terms computation of semantic weights, and term-to-term similarities; (b) construction of semantic vectors, and vector-to-vector similarities.

The first step consists of assigning POS tags to the lexical definitions a and b, which is a complex natural language processing task, carried out offline [290]. Similarly, the determination of semantic weights, computed through IDF, has a large spatial complexity, equivalent to a full scan of a text corpus ($O(C_t)$, where C_t is the number of terms in corpus C). However, this task is also typically performed offline, and the weights can be easily indexed. Therefore, the construction of semantic vectors can be considered to have a linear complexity $O(|\vec{a}| + |\vec{b}|) = O(n)$.

The computational complexity of the term-to-term similarity scores depends on the selected sim_t. WordNet sim_t measures vary from shortest-path algorithms (*path*, *wup*, *res*, etc.), to very complex lexical chains (*hso*). In the evaluation of our approach, we discuss the empirical time complexity of the ten measures included in the study (see Section 5.6.2). As a representative complexity for this step, we consider Dijkstra's classic shortest-path algorithm $O(|\vec{a}| \cdot |\vec{b}| \cdot (|W_E| + |W_V| log |W_V|)) = O(n^3)$, where W_E and W_V are the edges and vertices of WordNet. Hence, all the sim_t scores have to be pre-computed offline. This is the most computationally expensive step, but is manageable because the semantic vectors tend to be quite small (average $|a| \approx 48$).

Finally, the vector-to-vector measures need to construct a $|\vec{a}| \times |\vec{b}|$ similarity matrix [54, 75]. Given that sim_t are symmetric, the complexity of this step is $O(\frac{|\vec{a}| \times |\vec{b}|}{2}) = O(n^2)$. The overall upper bound of the lexical approach complexity is cubic: $O(n + n^3 + n^2) \leq O(n^3)$.

Considering the limited size of the OSM Semantic Network and its definitions (about 6,500 concepts), and applying the appropriate pre-computations, this cubic complexity does not represent a problem in this domain. If applied to OSM instances, as opposed to geographic concepts, radical optimisation

would certainly be needed. This approach to lexical similarity is evaluated in Sections 5.6 and 6.4. Its cognitive plausibility is obtained by observing the correlation with a human similarity dataset of geographic concepts, the MDSM evaluation dataset and our GeReSiD. The next section describes two techniques to combine the two components of NetLexSiM, the network and lexical similarity, into a hybrid measure.

3.8 NetLexSiM: hybrid similarity

Our approach to semantic similarity of OSM concepts, NetLexSiM, rests on two pillars: network and lexical similarity. The network similarity is based on the topological structure of the semantic network, i.e. the link patterns between concepts (Section 3.6), while the lexical similarity focuses on the natural language that users employ to describe the geographic concepts (Section 3.7). These two perspectives on concept similarity are not mutually exclusive, and should be considered as complementary.

The limitations of two computational approaches to the same problem can be overcome by combining them into an appropriate hybrid measure [45]. In the OSM Semantic Network, for example, some concepts might be situated in a densely connected area of the network, while having a sketchy lexical definition (e.g. 'stubs' in the wiki jargon). Similarly, other concepts are poorly linked, but have more exhaustive lexical definitions. When multiple measures are available, a representative average has to be extracted. The idea of combining multiple similarity measures will be further explored through the analogy of a 'similarity jury,' applied to lexical similarity measures in Section 5.7.

Considering two geographic concepts a and b, we have defined a network similarity measure $s_{net}(a,b)$, and a lexical similarity measure $s_{lex}(a,b)$. Both measures quantify the concept similarity with a real number in the interval $\Re \in [0,1]$, where 0 means minimum similarity, and 1 maximum similarity. In order to obtain a combined measure of similarity $s_{comb}(a,b)$, we define two combinations: *score combination* (s_{sc}), and *rank combination* (s_{rk}).

The score combination s_{sc} consists of the linear score combination of network and lexical similarities, weighted by a combination factor α:

$$s_{sc}(a,b) = \frac{\alpha \cdot s_{net}(a,b) + (1-\alpha) \cdot s_{lex}(a,b)}{2}, \quad \alpha \in [0,1] \tag{3.18}$$

The rank combination s_{rk}, on the other hand, is the linear combination of the pair rankings, normalised on the cardinality of the pair set:

$$\begin{aligned} rk_{comb}(a,b) &= \alpha \cdot rk(s_{net}(a,b)) + (1-\alpha) \cdot rk(s_{lex}(a,b)) \\ s_{rk}(a,b) &= \frac{|P| - rk_{comb}(a,b)}{|P|-1}, \quad rk_{comb} \in [1,|P|], \;\; s_{rk} \in [0,1] \end{aligned} \tag{3.19}$$

where rk is a ranking function, P a set of concept pairs, and α is the combination factor. While s_{sc} is a continuous function, s_{rk} is discrete. For example, in a set P of ten pairs, a pair of concepts (a, b) can have $s_{net} = .7$, resulting in $rk(s_{net}) = 3$ in the pair set. The lexical score $s_{lex} = .45$ might correspond to $rk(s_{lex}) = 8$. Fixing the value of α to .5, the score combination is $s_{sc} = .57$. The rank combination amounts to $rk_{comb} = 5.5$, therefore $s_{rk} = .5$.

The effectiveness of these two methods to combine network and lexical similarity is evaluated empirically in Section 6.5, showing the impact of parameter α on the cognitive plausibility of the resulting hybrid measure. This hybrid approach completes our similarity measure NetLexSiM, and provides a viable technique to quantify the semantic similarity of OSM concepts, tapping the OSM Semantic Network. The next section presents an alternative approach to similarity, which does not focus on individual concepts or features, but on map viewports.

3.9 Holistic viewport similarity

In the previous sections, we focused on techniques to compute the semantic similarity of geographic concepts in the OSM Semantic Network. When using GIR systems, users are presented with geographic data displayed in 'viewports,' defined as rectangular, bi-dimensional images rendered on a screen. While NetLexSiM computes the semantic similarity of individual concepts, it is possible to compare semantically *entire viewports*, in a holistic way. Hence, we devised a viewport-based GIR system, by examining the structure of a typical web map, and constructing vector-based semantic descriptors for viewports.

In our system, the user has the possibility of submitting a viewport to the system, which exploits the semantic descriptors to retrieve semantically similar viewports, enabling an alternative exploration of geographic data. Although this approach was developed in isolation as an initial proof of concept, it will certainly benefit from an integration with NetLexSiM. Using the OSM Semantic Network and NetLexSiM to add a fine-grained measure of conceptual similarity will increase the ability of the approach to capture holistic viewport semantics, obtaining high cognitive plausibility. This work was published in [23].

This section is organised as follows. First, we clarify the concept of map viewports (Section 3.9.1), and subsequently we devise semantic descriptors (Section 3.9.2). Given the very large size of the viewport space, sampling techniques are discussed (Section 3.9.3). The semantic content of the viewports is finally compared using similarity measures (Section 3.9.4).

3.9.1 Viewports

Our approach to GIR focuses on map *viewports*, treated as semantic units. In a session in our GIR system, the user can retrieve viewports that are similar

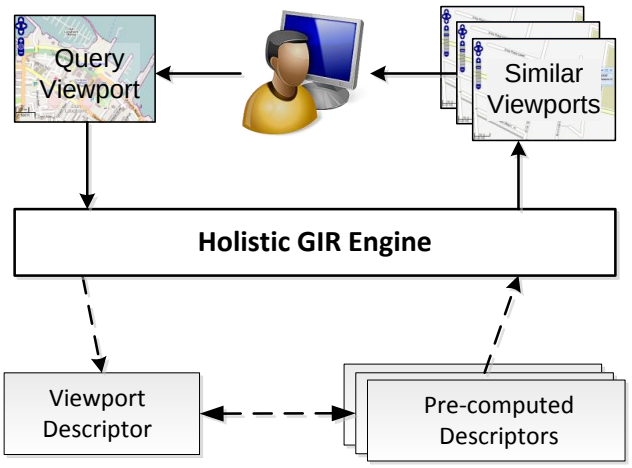

Figure 3.9: The architecture of a viewport-based, holistic GIR system. The user submits a query viewport to the system, and the system retrieves the most similar viewports from a set of precomputed semantic descriptors.

Zoom Level	Meters per pixel	Scale	Visible Types	Description
1	78,271	$1:223M$	2	Region
3	19,568	$1:55M$	2	–
5	4,892	$1:14M$	3	Country
7	1,123	$1:3.5M$	5	–
9	306	$1:871K$	10	County
11	76	$1:217K$	23	–
13	19	$1:54K$	44	Neighbourhood
15	5	$1:13K$	64	–
17	1	1 : 3400	93	Building

Table 3.5: Overview of the zoom levels of a CloudMade web map. Total number of feature types: 101.

to a query viewport, indicated as fulfilling the users' spatial information need. To capture the semantic content of a viewport, it is useful to start from the visualisation structure of a typical web map. The system compares the query viewport with pre-computed viewports, and returns to the user the most similar viewports it has found. This GIR architecture is schematised in Figure 3.9. In order to test our approach, we utilised the OpenStreetMap vector dataset, released under a Creative Commons license [113]. As a representative of typical web mapping, we selected the CloudMade service, which renders OpenStreetMap data as interactive online maps.[20] As opposed to other commercial geo-services, this service enables exploration of the internal structure of a viewport, and its underlying geographic content.

In the CloudMade maps, the scale can be set to 19 discrete zoom levels, ranging from scale 1:446M (zoom level 0) to 1:1700 (zoom level 18). The map scale is controlled by Equation 3.20, where y is either the distance in meters or the map scale, and z is the zoom level. The constant C is $78,271$ in the case of

[20] `http://maps.cloudmade.com`

meters per pixels, and $223 \cdot 10^6$ in the case of map scale. This equation allows the conversion between map scale, screen pixels, and zoom levels:

$$y = C\, 2^{1-z} \qquad z \in [0, 18] \tag{3.20}$$

At each zoom level, the map displays certain types of features, e.g. at the region level, only countries and seas are shown. For the purpose of our study, the geographic dataset has been subdivided into 101 feature types, closely modelled on the visualisation rules of the CloudMadeMap.[21] For example, types include 'restaurant,' 'stadium,' and 'prison.' The visibility of each type per zoom level is defined as a range, e.g. restaurants are visible when $z \in [16, 18]$, while stadiums, being generally larger objects, in range $z \in [14, 18]$. For the sake of clarity, all the notations used in this section are displayed in Table 3.6. Intuitively, the number of visible types increases as the scale decreases. The characteristics of each zoom level are summarised in Table 3.5.

In this context, a viewport $v_{bb,z}$ is defined by a bounding box bb, specified by the latitude/longitude coordinates of its bottom-left and top-right corners, and a zoom level $z \in [0, 18]$ as defined in Table 3.5. As stated in Section 2.5.6, a viewport can be seen as the visualisation unit of geographic data in a web map. A user session on a web map consists of a sequence of manipulative actions on the map, such as panning and zooming, resulting in the visualisation of a sequence of viewports $\{v_1 \dots v_n\}$. In our GIR system, the user can select a viewport by drawing a bounding box on the map, and the selected viewport v_s is used as a query, described in the next section.

3.9.2 Holistic viewport descriptors

In order to retrieve semantically similar viewports, our GIR system constructs a holistic semantic descriptor for each viewport v in a vector space model. To extract the overall semantic content from a viewport, the system performs spatial queries on the OpenStreetMap dataset. Given the input viewport v, the system will perform spatial queries to retrieve F_v, all the visible features in that viewport (Equation 3.21):

$$\forall t \in S_z, \quad q(bb, z, t) \rightarrow F_t, \qquad F_v = \{F_{t_1} \dots F_{t_{|S_z|}}\} \tag{3.21}$$

The service can now compute a holistic descriptor D_v, combining all the visible types S_z in a multidimensional vector as in Equation 3.22, where n is the cardinality $|S_z|$, $t \in S_z$, and w are non-negative normalised weights.

$$D_v = w_1 t_1 + w_2 t_2 + \dots + w_{n-1} t_{n-1} + w_n t_n, \qquad w \in [0, 1], \quad \sum_{i=1}^{n} w_i = 1 \tag{3.22}$$

In order to characterise D_v, we propose four ways to compute the weights w: linear, logarithmic, information-theoretic, and surface-based approaches.

[21] The visibility rules are defined at `http://maps.cloudmade.com/editor`

Symbol	Description
z	Zoom level $\in [0, 18]$ (see Table 3.20 for details).
bb	Bounding box, specified by the latitude/longitude coordinates of its bottom-left and top-right corners.
v	A viewport in bounding box bb at zoom level z. h_v and w_v is the viewport height and width in screen pixels.
t	A type of map feature (e.g. 'restaurant,' 'prison,' etc.). In this work 101 types were defined.
$\psi(z)$	Function mapping the visibility of feature types at zoom level z. $S_z \leftarrow \psi(z)$.
S_z	Set of visible t at zoom level z. $S_z = \{t_0 \dots t_n\}$, where $\forall t$ is visible at zoom level z.
D_v	Semantic descriptor of viewport v.
$q(bb, z, t)$	Spatial query on bounding box bb, zoom level z, and feature type t. $q(bb, z, t) \rightarrow F_t$
F	Set of all existing map features.
F_t	Set of features of type t, $F_t = \{f_0 \dots f_n\}$
F_v	Set of features visible in viewport v. $F_v = \{F_{t_1} \dots F_{t_n}\}$
$I(t)$	Self-information of type t, assuming a random distribution of types in the map.
$a(f)$	Area of feature f.
V_g	Set of viewports extracted from a geographic area g.

Table 3.6: Notations for viewport similarity

(a) Linear weights. The simplest approach consists of assigning them proportionally to the cardinality of sets $F_t \in F_v$, using the normalised cardinality:

$$w_i = \frac{|F_{t_i}|}{\sum_{j=1}^{n} |F_{t_j}|} \tag{3.23}$$

The main limitation of this approach lies in the fact that the statistical distribution of types t is heavily skewed in favour of very frequent features, such as 'road.' In a viewport of an urban area, the number of features 'road' is often greater than other types by several orders of magnitude, such as 'restaurants,' which matches our intuition about the fact that roads are very common map objects, while restaurants are less frequent. For example, it is not uncommon to find 2 restaurants, and 200 roads in a viewport. In this case, the weighting function in Equation 3.23 would assign a very low weight to the type 'restaurant,' and an extremely high weight to 'road.'

(b) Logarithmic weights. A variant that focuses on magnitude rather than number of features is the following, where δ is a positive quantity that prevents the nullification of a term if $|F_{t_i}| = 1$:

$$w_i = \frac{log(|F_{t_i}| + \delta)}{\sum_{j=1}^{n} log(|F_{t_j}| + \delta)}, \qquad \delta = 1 \tag{3.24}$$

This logarithmic version is less sensitive to small changes in the statistical distribution of types t, and tends to preserve the importance of infrequent features.

(c) Information theoretical weights. A second variant to compute weights w_i taking into account the statistical occurrence of feature types, is based on the information theoretical approach [265]. Given a set of spatial features F, the probability p and the self-information I of feature type t randomly from the dataset are defined as in Equation 3.25. The self-information of type t can then be used to weight its importance in the vector, by combining it with the number of features:

$$p(t) = \frac{|F_t|}{|F|} \qquad I(t) = -log(p(t)) \qquad w_i = \frac{I(t_i)|F_{t_i}|}{\sum_{j=1}^{n} I(t_j)|F_{t_j}|} \tag{3.25}$$

In this case, the importance of a type t in the descriptor D_v is increased or reduced depending on its frequency in the dataset. Therefore features of type 'road' carry low self-information, while features 'restaurant' are emphasised. Although this weighting approach intuitively seems the most promising among the three we have presented (Equations 3.23, 3.24, and 3.25), it can assign very high weights to rare features. While in some cases this might be a desirable behaviour (for example to detect landmarks), in general it risks skewing the descriptor towards unusual features, regardless of their actual semantic weight in the viewport.

(d) Area weights. With features modelled as a polygon, it is possible to attribute a weight proportionally to the feature area, on the assumption that large features should have higher semantic importance in the descriptor. Defining the feature area as $a(f)$, and $a(F_t)$ as the sum of all the areas of the features in the set, the surface weights are computed as:

$$w_i = \frac{a(F_{t_i})}{\sum_{j=1}^{n} a(F_{t_j})} \qquad a(F_t) = \sum a(f), \quad \forall f \in F_t \tag{3.26}$$

In this approach, the feature area is weighted against the area of the other features, and not of the viewport. Thus, the weight can account for overlapping features.

The effectiveness of these four approaches to weight the semantic types in the viewport descriptor, summarised in Table 3.7, largely depends on the specific application context. As each approach captures different aspects of the holistic semantics of a viewport, and presents specific limitations, the weights can be computed by averaging different approaches. Without doubt, one of the main advantages of such vector-based viewport semantic descriptors is

Approach	Description
(a) Linear	Number of features. When types have different magnitude, large magnitudes take a large section of the descriptor.
(b) Logarithmic	Log of number of features. Represents the magnitude of types. Low sensitivity when types have the same magnitude.
(c) Self-Information	Self-Information of feature type. Common features have low weight, while unusual feature types have very high weight.
(d) Area	Sum of feature areas. Large features have high weight. Not computable when feature is not a polygon (e.g. for points of interest)

Table 3.7: Weighting approaches in semantic viewport descriptor D_v

the wide range of techniques to compare, cluster, and classify them, discussed in the next section.

3.9.3 Sampling the Viewport Space

In order to describe the semantics of a web map viewport, we have defined a vector-based descriptor D_v, in four variants (linear, algorithmic, self-information, and area). Given a web map covering a certain geographic area g, e.g. Ireland, we aim at extracting a number of descriptors that represent its semantics. The viewport space is the set of all possible viewports in g at zoom level z. The area g can be sampled in a number of viewports $V_g = \{v_1 \dots v_n\}$, where n is the desired number of viewports. The theoretical number of viewports that can be extracted in a geographic area delimited by a bounding box bb_g at zoom level z, where h_g, w_g are the height and width of the geographic area converted into pixels with Equation 3.20. h_v and w_v are the height and width of the viewport in pixels:

$$|V_g| = (h_g - h_v)(w_g - w_v) \tag{3.27}$$

For example, a geographic area g delimited by a bounding box of size $\approx 300 \times 230\ km^2$ corresponds at $z = 9$ (county level) to a screen of 4096×3072 pixels. Sampling g with 1024×768 pixel viewports, a common resolution for web maps, the possible viewports amount to ≈ 7.6 million. With higher zoom levels, the number of possible viewports increase rapidly following the power law in Equation 3.20. It is therefore evident that a sampling technique has to be utilised to extract a computable number of viewports, in particular for high zoom levels.

The most intuitive way of choosing viewports is based on user interests. Viewports in which user activity is performed are automatically included in the sample. However, to overcome the cold start problem that arises in this case, a general sampling mechanism is necessary to index g. A possible technique is that of random sampling, based on the law of large numbers. Even though it is difficult to compute all the possible viewports, it is possible to

extract a sufficient number of random viewports to represent accurately the whole set of viewports, setting the sample size to a limit β, as shown in Equation 3.28. In order to determine β, a confidence level and a confidence interval have to be chosen.

$$|V_g|_\beta = \frac{\beta}{|V_g|}(h_g - h_v)(w_g - w_v) \qquad 0 < \beta < |V_g| \tag{3.28}$$

Similarly, it is possible to sample g by defining an arbitrary grid γ of pixels h_γ and w_γ, which reduces the number of viewports as follows:

$$|V_g|_\gamma = \frac{(h_g - h_v)(w_g - w_v)}{h_\gamma w_\gamma} \qquad 0 < h_\gamma < h_g, \;\; 0 < w_\gamma < w_g \tag{3.29}$$

A third possibility is a combination of grid and random sampling. The geographic area g is divided into an arbitrary number of grid cells, and each cell is sampled randomly. The choice of the sampling technique (random, grid-based or both) has to be done on an empirical basis, depending on the specific application context. Once a sample V_g has been selected, the corresponding descriptors D_v can be computed. Subsequently, the system can compare, cluster, and retrieve viewports (see Figure 3.9). The next section describes comparison techniques for the semantic descriptors.

3.9.4 Comparing Viewport Descriptors

A geographic area g can sampled as a set of viewports V_g, with the techniques described in the previous section. The corresponding descriptors D_v can then be pre-computed through one of the approaches presented in Section 3.9.2. The similarity of two viewports is therefore the similarity of their descriptors (Equation 3.30).

$$s(v_a, v_b) = s(D_{v_a}, D_{v_b}) \quad s(v_a, v_b) = s(v_b, v_a) \quad 0 \leq s(v_a, v_b) \leq 1 \tag{3.30}$$

Given that these descriptors are multidimensional vectors, encoding semantic aspects of the viewports, it is possible to compare them with well-known techniques in a vector space [255, 138, 294]. In the viewport semantic vector space, every viewport can be modelled as a row in a multidimensional matrix $|V_g| \times |t|$, having a column for each feature type t. Vector similarity can be computed with the Euclidean, cosine, Chebyshev, and Manhattan distance [177].

In a semantic information retrieval system in which the user submits a viewport v_q as a query, the most similar k viewports must be retrieved and displayed to them. To do so, these distance measures can be efficiently computed between the descriptor of the query viewport D_{v_q} and all the pre-computed descriptors D_v, where $v \in V_g$. Additionally, the similarity computation can be

constrained in several ways to match specific user information needs. Among others, common constraints are the maximum or minimum distance from the query viewport v_q, a zoom level range ($z \in [z_{min}, z_{max}]$), and a weight constraint on a feature type ($w_{min} < w_t < w_{max}$).

A preliminary evaluation to this approach is reported in Section 6.6, outlining a case study, in which the descriptors are used to capture the holistic semantics of viewports taken from the Dublin area in Ireland, and are used to retrieve semantically similar viewports.

3.10 Summary

In this chapter, we have outlined the approaches that constitute our contribution to the support of the semantics of OSM, in the emerging context of VGI. Focusing on the OSM vector dataset, we have developed an alignment technique with DBpedia, a novel semantic network, and NetLexSiM, a hybrid approach to computing concept-to-concept similarity.

The chapter started by outlining our preliminary work on map personalisation and implicit feedback, which indicated the semantic issues with the OSM vector dataset (Section 3.2). In this context, the development of RecoMap, a system that relies on implicit feedback to recommend spatial items, highlighted directions for supporting the OSM model. Subsequently, we devised an approach to integrating the OSM dataset with open knowledge bases, expanding its semantics with DBpedia.

In order to provide a richer semantic conceptualisation for OSM, a novel semantic network was devised, the OSM Semantic Network, extracted from the OSM Wiki website (Section 3.3). This resource is designed to provide semantic support for OSM (Section 3.4). Computing the semantic relatedness and similarity of the concepts represented in this network has many promising applications in Geographic Information Systems (GIS), including data mining, map personalisation, and GIR (Section 3.5).

To devise NetLexSiM, our similarity measure for OSM concepts, two aspects were considered. First, the network-based semantic similarity, using co-citation measures (Section 3.6). Second, a natural language processing technique to model the lexical similarity of concept definitions (Section 3.7). These two perspectives are combined in a hybrid measure (Section 3.8). Furthermore, we have explored an alternative approach to semantic similarity in Web maps, going beyond the traditional concept-to-concept comparison. Our GIR approach considers viewports to be basic, holistic semantic units, which can be compared semantically to a query viewport (Section 3.9).

Details of the implementation of these techniques are outlined in the next chapter. The body of empirical evidence that we have collected to validate these approaches will be discussed in Chapters 5 and 6.

CHAPTER

FOUR

IMPLEMENTATION

4.1 Overview

To contribute to the semantics of OpenStreetMap (OSM), we have developed a novel semantic resource, the OSM Semantic Network. This chapter details the implementation of the tool that constructs the semantic network, and describes the preliminary work that highlighted the semantic gap in Volunteered Geographic Information (VGI). As part of that preliminary work, we designed a Web platform for map personalisation and visualisation, and we conducted an online survey of the open source technologies utilised in the platform. This initial phase also included the recommender system RecoMap, covered in Chapter 3.

In the landscape of evolving technologies for spatial crowdsourcing projects, we have put particular emphasis on free and open-source software (FOSS). Open source tools are a fundamental pillar in VGI and collaborative, non-profit endeavours. The infrastructure that makes VGI possible includes web hosting services, open source map editors, forums, and wiki websites. Contributors produce several open geospatial datasets, relying on open source tools to collect, validate, and disseminate crowd-sourced geo-data, breaking the constraints of traditional vertical production modes.

This chapter is organised as follows. Section 4.2 describes the OSM Wiki Crawler, which we designed to extract the OSM Semantic Network. This tool constructs a novel semantic resource, the OSM Semantic Network, representing concepts utilised in OSM. The crawler extracts the semantic network from a dedicated website, the OSM Wiki website, where contributors negotiate a common conceptual lexicon.[1]

The semantic gap in OSM was noticed in the preliminary research that we carried out in the area of spatial recommender systems, map personalisation, and implicit feedback. This work was conducted in collaboration with the National University of Ireland, Maynooth, in the framework of the cluster Strategic Research in Advanced Geotechnologies (StratAG).[2] While investigating the state-of-the-art technological landscape to identify suitable technologies for a Web platform for map personalisation and visualisation, we have

[1] `http://github.com/ucd-spatial/OsmWikiCrawler`

[2] `http://www.nuim.ie`, `http://www.stratag.ie`

conducted a user survey, asking open source developers their opinion about a set of tools for the geo-Web, including Web application frameworks, spatial DBMSs, and toolkits for Web GUI, with the purpose of identifying the optimal technologies for our context. Users have consistently expressed satisfaction with the degree of stability and standardisation of these tools, and less so for user-friendliness and documentation (Section 4.3). Our Web platform for map personalisation and visualisation was then developed to host Web services tailored to exploit implicit feedback mechanisms, with the purpose of providing a range of personalisation services (Section 4.4).

4.2 OpenStreetMap Wiki Crawler

This section describes in detail the implementation of the OSM Wiki Crawler, the open source tool that we have developed to extract the OSM Semantic Network (see Section 3.3). The OSM Wiki Crawler is released under the GPLv3 license, and is freely available online.[3]

The OSM Semantic Network is a novel semantic network, whose vertices represent OSM geographic concepts, and whose edges model relationships between concepts, resulting in a directed graph. In order to extract the directed graph **G** from the OSM Wiki website,[4] we implemented the OSM Wiki Crawler, an open source tool tailored to the OSM Wiki content structure. The purpose of this semantic crawler is the extraction of a semantic network from a dynamic and complex wiki website, encoding geographic knowledge that can be utilised for various tasks – in this paper we focus on the computation of semantic similarity. Although the crawler focuses on the OSM Wiki website, its general approach can be adopted to extract a semantic network from wiki, open content websites. Among many possible applications, we have used the OSM Semantic Network to compute the semantic similarity of geo-concepts (see Sections 3.6 and 3.6).

The extracted network is stored in Resource Description Framework (RDF), containing a set of statements of the format $\langle subject, predicate, object \rangle$, logically equivalent to a labelled, directed graph.[5] The crawler starts from cluster pages and scans their hyperlinks, creating RDF statements until the queue of pages to scan is empty (see Algorithm 1). In order to make sure that the crawler covers the entire website, we have included as a cluster the page `osmwiki:Special:AllPages`, which contains links to all of the pages of the website.

For each page scanned, the crawler extracts the following information, if available: OSM tag descriptions, internal links, and links to Wikipedia pages. A heuristic function assigns OSM tags to the equivalent terms in the LinkedGeoData ontology [15]. The heuristic is based on lexical matching between the OSM tag and the LinkedGeoData term. For example, the

[3] `http://github.com/ucd-spatial/OsmWikiCrawler`
[4] `http://wiki.openstreetmap.org`
[5] `http://www.w3.org/RDF`

Algorithm 1: The OSM Wiki crawler

input : set P_c of cluster pages
σ = maximum length for key/value strings (e.g. 30 characters)
output: set S of $\langle subject, predicate, object \rangle$ statements, equivalent to directed graph **G**

```
1  S ← ∅,  queue ← ∅,  visited ← ∅
2  foreach url ∈ P_c do push url to queue
3  while queue ≠ ∅ do
4      url ← pop head from queue
5      if url ∈ visited then skip
6      if url matches 'Tag:key=value' or 'Key:key' then
7          extract key/value
8          if key/value are longer than σ then
9              pseudo-tag detected, skip
10         if key/value contain invalid characters then
11             pseudo-tag detected, skip
12         add ⟨url, is_a, class⟩ to S
13         add ⟨url, keyLabel, key⟩ to S
14         add ⟨url, valueLabel, value⟩ to S
15         if description found in page(url) for key/value then
16             if ⟨url, rdf_comment, ∗⟩ ∈ S then
17                 merge descriptions
18             else
19                 add ⟨url, rdf_comment, description⟩ to S
20             if key/value or value matches term lgd in LinkedGeoData then
21                 add ⟨url, equivalent_class, lgd⟩ to S
22     else
23         add ⟨url, is_a, cluster⟩ to S
24     add url to visited
25     H ← hyperlinks to OSM wiki or Wikipedia in page(url)
26     foreach h ∈ H do
27         add ⟨url, link, h⟩ to S
28         if h ∉ visited then push h to queue
```

OSM tag *amenity=fountain*, is matched against `lgdo:AmenityFountain`.[6] If the *key=value* pair is not defined, only the value is considered (e.g. `lgdo:Fountain`). We have validated this approach by observing that, in a random sample of size 30, all the mappings to LinkedGeoData were correct. With P_c initialised to `osmwiki:Special:AllPages`, `osmwiki:Proposed_features`, and `osmwiki:Map_Features`, the crawler scanned 5,407 pages, generating 34,115 RDF statements (extracted on February 1, 2012). The detailed structure of the resulting semantic network was described in Section 3.3.

The strategy adopted by the crawler to extract geographic concepts from raw HTML data is outlined in Algorithm 1. An important aspect of the extraction process is the detection and removal of noise. The quality of the OSM Semantic Network largely depends on the precision and recall of this automatic extraction. The automatic identification of OSM tags and their semantic

[6] 'lgdo:' stands for the namespace `http://linkedgeodata.org/ontology/`

relations in semi-structured wiki text is challenging [14]. The crawler relies on regular expressions to find key/value pairs, e.g. `(\w+)=(\w+):(.+)`. This simple approach tends to obtain a large number of false positives, resulting in invalid tags. Some cases are too short (1-2 character-long tags), while others are too long (30+ character-long tags). Hence, to reduce this noise, we introduced a threshold σ, which indicates the maximum length of tag keys and values. Empirically, we found that discarding matches shorter than 3 characters, and setting σ to 30 characters, increases the precision of matches, without a significant decrease in recall.

The complete source code of the tool, written in the Groovy language,[7] including detailed instructions on how to run it, is available online, including pre-extracted RDF networks.[8] The OSM Semantic Network was conceived in the context of preliminary work on map personalisation and implicit feedback. The next section presents the Web platform that we developed to host geo-personalisation services.

4.3 Survey of open source Web mapping tools

In the process of setting up our Web platform for map personalisation and visualisation, described in Section 4.4, we have investigated the area of open source geospatial software, with a particular focus on Web mapping technologies. The purpose of this survey was to identify the best technologies to use, as well as to provide 'grounded' recommendations to other developers who are operating in the area of web mapping. The survey was conducted in 2011, in the framework of Strategic Research in Advanced Geotechnologies (StratAG), in collaboration with the National University of Ireland, Maynooth.[9] The results of this survey have been published in [21].

The production and usage of FOSS have grown considerably over the past decade, mainly due to its improved quality and economic factors [59]. FOSS combines the notion of free software and open source software. 'Free software' relates to the user's freedom to run, copy, distribute, study, change and improve the software [277]. 'Open source' software is generally distributed under licences approved by the Open Source Initiative (OSI). Within the open source ecosystem, Web technologies have emerged as one of the dominant domains. This has occurred due to the advent of Web 2.0 in the last decade, which required free, interoperable, standard-compliant software components. As a result, several open source projects challenged commercial competitors and reached a considerable user base.

In parallel, desktop-based Geographic Information Systems (GIS) have successfully merged with Web 2.0 technologies resulting in Web mapping, which has become a ubiquitous feature of online services. The availability of open geographical datasets from sources such as OpenStreetMap has fu-

[7] `http://groovy.codehaus.org`
[8] `http://github.com/ucd-spatial/OsmWikiCrawler`
[9] `http://www.nuim.ie`, `http://www.stratag.ie`

elled the development of a large number and variety of open source Web mapping projects, contributing an alternative to well-established commercial services like Google Maps. Within this context, the Open-Geospatial Consortium (OGC) has proposed standards for geographical data and facilitated the design of interoperable open source GIS tools. With such an abundance of different technologies available, it is difficult for a developer to weigh the merits of each approach. Although there is a body of theoretical work in this area, the speed of development has created a gap in the knowledge regarding the practical aspects of developing open source Web-based geospatial data delivery services. The next Section describes the specific technologies we have analysed.

4.3.1 Open source technologies

In order to make an informed decision regarding which technologies to adopt in our Web platform, information was collected from several Web communities and the academic literature. To reduce the knowledge gap regarding real-life experiences of software development with these projects, an online survey was devised to collect opinions from the communities of users and contributors, which constitute an essential part of the open source phenomenon.

Two projects for each of the seven categories were identified as suitable for developing a Web mapping architecture and these were studied further. In total 14 projects – *OpenLayers, Ext JS, Prototype, MooTools, GeoServer, MapServer, GeoTools, Java Topology Suite, Ruby on Rails, Grails, Hibernate, Hibernate Spatial, PostGIS*, and *MySQL* – were examined and included in the survey described in this article. The projects are not in direct competition, as they are often interdependent. While in some situations, a technology can be a clone of another (e.g. Ruby On Rails and Grails), in other cases it can be a more specialised spatial extension of another (e.g. Hibernate and Hibernate Spatial). Other projects which appear similar, may in fact be responsible for different functional areas and so are not competing with each other (e.g. OpenLayers and Ext JS). This survey does not attempt to cover all the open source projects currently active in the geospatial area but focuses on the specific domain of Web mapping. The online communities associated with these projects are one of the most interesting aspects of the open source software projects.

Table 4.1 reports the year the project started, the software license, well-known corporate and academic users and the link to the official project websites (accessed on February 24, 2011). Each project provides an official forum (or a mailing list) for general users and one devoted to the project developers and contributors. The number of users and developers presented in the table is based on the number of registered members on such forums. Certain forum administrators, such as those of MySQL, did not disclose the number of users. An overall description of the 14 projects is presented in the remainder of this section. Section 4.3.2 will subsequently present the methodology of the online questionnaire.

Project	Owner	Year	License	#Users	#Contrib	Website & Users
OpenLayers	(MetaCarta until 2007) Open Source Geospatial Foundation	2006	BSD-like	1,450	630	http://openlayers.org *MetaCarta*
Ext JS	Ext	2006	GPL/comm	127,000	*n/a*	http://www.extjs.com *Adobe, Amazon.com, Microsoft, Sony*
Prototype JS	Prototype Core Team, 37signals	2006	MIT	3,200	1,080	http://www.prototypejs.org *Apple, CNN.com, Nasa, Microsoft, Twitter, eBay*
MooTools	The MooTools Dev Team	2005	MIT	2,500	14	http://mootools.net *Jeep, Ferrari*
MapServer	University of Minnesota (UMN)	1996	X/MIT	1,880	228	http://mapserver.org *Camptocamp,WebMapIt,Syncera IT Solutions*
GeoServer	Lime Group, The Open Planning Project, Refractions Research	2001	GNU GPL	*n/a*	52	http://geoserver.org *Alkante, Camptocamp, GeoSolutions, Landgate, Great Lakes Commission*
GeoTools	Open Geospatial Consortium	1996	GNU LGPL	1,059	434	http://www.geotools.org *VisionOfBritain, Institut de Recherche pour le Développement*
JTS	Martin Davis	2000	GNU LGPL	*n/a*	77	http://www.vividsolutions.com/jts *GeoConnections, British Columbia Ministry of Sustainable Resource Management*
Ruby On Rails	David Heinemeier Hansson, 37signals	2004	MIT	20,000	1,500	http://rubyonrails.org *Basecamp, Twitter, Yellow Pages*
Grails	SpringSource (WMWare)	2006	Apache v2.0	4,600	890	http://grails.org *Sky.com, Wired.com*
Hibernate	Red Hat	2001	GNU LGPL	57,000	*n/a*	http://www.hibernate.org *UnionBank, Warner Music Group*
Hibernate Spatial	GeoVise	2003	GNU LGPL	100	5	http://www.hibernatespatial.org *N.A.*
PostGIS	Refractions Research	2001	GNU GPL	1,790	274	http://postgis.refractions.net *GlobeXplorer,Institut Gographique National, France*
MySQL	Sun Microsystems	1994	GNU GPL	*n/a*	*n/a*	http://www.mysql.com *Twitter, LinkedIn, United Nations FAO*

Table 4.1: Open source projects: owner, year of foundation, estimate of number of users and contributors (URLs accessed on February 24, 2011)

Web GUI Libraries

Thanks to the diffusion of AJAX-oriented Web technologies, Web GIS has experienced remarkable growth over the past decade. Well-known services such as Google Maps and Virtual Earth provide developers with powerful and free Web tools. Although these products are very popular, they are fully controlled by companies that can decide to alter or discontinue them arbitrarily. Furthermore, there is no clear separation between the visualisation tools and data, the two are mixed in a seamless way and so the GUI cannot be customised. On the contrary, open source mapping libraries offer viable alternatives to build powerful interactive Web maps, allowing the developer to choose the data sources and formats. In this survey OpenLayers, a popular open source Web mapping project, and Ext JS, a library to build complex Web user interfaces, are assessed.

OpenLayers is a JavaScript library for displaying spatial data in web browsers, without server-side dependencies. OpenLayers implements a JavaScript API for building rich web-based geographic applications, constituting the main Open-Source alternative to the equivalent commercial tools. As with many open source geospatial technologies, OpenLayers implements industry-standard methods for geographic data access, such as the OGC Web Map Service and Web Feature Service standards. The architecture of the framework is written in object-oriented JavaScript, using components from Prototype. At the conceptual level, OpenLayers aims to separate spatial visualisation and manipulation tools from spatial data, an idea that is not implemented in the commercial equivalent. OpenLayers is also widely used to display OpenStreetMap data.[10]

Ext JS is a cross-browser JavaScript library for building Web interfaces. It includes a set of User Interface widgets, an extensible component model and an API. The Ext JS library and its related products have experienced success recently, becoming ubiquitous in complex Web applications. The library is currently used by several major international corporations for Intranet and Internet websites. Ext JS can work as a stand-alone library and can integrate with the Prototype Javascript library. The work has been extended by a project called GeoExt which embeds OpenLayers with Ext JS to provide a framework to build desktop-like Web GIS applications.[11]

AJAX Libraries

Before the advent of AJAX, interaction with Web pages was a synchronous cycle of request/response between the client and the server. AJAX, meaning *Asynchronous JavaScript + XML*, is not a specific technology but an approach that can be implemented with diverse technologies. Such an approach introduced an intermediate layer between client and server, letting the client request data for a specific subsection of the content, adding asynchronous calls between the atomic synchronous calls.

[10] `http://www.openstreetmap.org`
[11] `http://www.geoext.org`

Although the core idea was first presented in 1992, it was not until Google released Gmail and Google Maps in 2005 that AJAX went mainstream. This approach adopted in Google Maps opened several possibilities to the Web GIS field, which, thanks to AJAX, has grown considerably. In this survey we have selected two of the most popular Javascript libraries that implement the principles of AJAX to address the diverse needs of Web development.

Prototype is a framework that provides functionality to assist JavaScript development. It provides several features, ranging from programming shortcuts to functions for dealing with AJAX requests. The framework extends object-orientation in Javascript and introduces a class-based design. In contrast to other JavaScript frameworks like jQuery, Prototype extends the W3C DOM, which is controversial and debated in the Web developer community and will probably undergo a major revision. Prototype is currently being used in a number of open source and commercial projects to implement AJAX patterns and build JavaScript classes. The library is prominent on the Web, where it is used by Apple, Microsoft and other major groups in their Web applications.

MooTools (My Object-Oriented Tools) is also a web-application framework to support JavaScript development. MooTools is based on the traditional object-oriented programming paradigm, with a similar class and inheritance structure as Java. This makes the framework intuitive to programmers with an object-oriented background. MooTools emphasises modularity and code reuse which can be customised depending on the application being developed. There are many advantages of using MooTools: of particular note are the built-in methods to handle Ajax requests, CSS and DOM elements. In addition, at present, MooTools works independently of specific browsers and operates across an array of popular platforms. It is used by several corporations, including Jeep and Ferrari for their web applications. Like Prototype, MooTools permits the easy integration of interactivity into a Web page.

Web Mapping Servers

Web mapping servers act as a framework to publish a GIS application online. Generally these servers include functionality to query spatial DBMS, projection support, integration with other geographic libraries as well as vector and raster support. Moreover, interoperable Web standards have been developed by the OGC in support of Web mapping, including Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), Geography Markup Language (GML), and Styled Layer Descriptor (SLD). WMS and WFS are widely used services to request maps and get information about the geographic features. These OGC standards have allowed numerous projects to be published online since the beginning of the FOSS movement. MapServer and GeoServer have been included in this survey.

MapServer is a widely used open source Web mapping project for developing interactive and interoperable Web GIS applications. Fully compliant with various OGC standards such as WMS, WFS, WCS and SLD, MapServer

supports an extensive variety of vector and raster formats. MapServer runs as a CGI (Common Gateway Interface) application with Apache and Microsoft Web servers. MapServer has strong cartographic support and dynamic capabilities. A *Map File* defines styles and symbology that are used to render a fully customised map. MapServer also has built-in support for many scripting languages such as Perl, Java, .NET, PHP and direct databases support for PostGIS, Oracle Spatial and MySQL.

GeoServer, developed more recently than MapServer, is a popular open source GIS project. It has a web administration tool to configure its spatial and non-spatial options. Apart from WMS, WCS and WFS, it additionally supports the editing of feature services on the client side using WFS-Transactional. GeoServer is built on Java technology and runs on an integrated Jetty Web Server. It relies heavily on GeoTools. A significant feature of GeoServer is the use of SLD, an OGC standard to render the visual style of the map. Other features include an integrated OpenLayers, Google Earth support for overlays using Keyhole Markup Language (KML), GeoWebCache for tile mapping and wide support for many DBMS like PostGIS, ArcSDE, Oracle and DB2.

Spatial Libraries

Spatial libraries provide spatial software components which can be exploited in desktop and Web GIS applications. There are several open source spatial libraries available online. Proj.4, an open source cartographic projection library is one such example that is widely used by many client and server side GIS applications, including MapServer. Other examples such as Geometry Engine-Open Source (GEOS) define the topological predicates and operations required to render a map. This survey covers GeoTools and the Java Topology Suite (JTS).

GeoTools is an open source Java library that provides advanced GIS functionalities. It supports vector and raster geospatial data formats, DBMS access and rendering of complex maps. Being one of the oldest projects of Open Geospatial Foundation, it is fully compliant with OGC specifications including GML, WMS, WFS, Grid Coverage, Coordinate Transformation and SLD. The JTS can be used in conjunction with GeoTools as a geometry model for vector features. This library can be used in client and server-side geospatial applications. It is currently used by GeoServer for a range of geographic tasks.

Java Topology Suite (JTS) is a widely used open source Java library for handling 2D geometrical operations. It conforms to the geometry model and API defined in the Simple Features Specification (SFS) proposed by OGC.[12] JTS provides access to simple functions such as buffering, overlays (intersection, union and difference) as well as more advanced functionalities like spatial algorithms and structure support (spatial indexing, planar graph framework, geometry simplification, Delaunay triangulation, and precision reduction). This library is used in a range of open source geospatial projects including GeoTools, PostGIS, GeoServer and Hibernate Spatial.

[12] `http://www.opengeospatial.org/standards/sfa`

Web Application Frameworks

As part of the Web 2.0 transition, described by O'Reilly [226], software applications have been moving to the World Wide Web. Developing powerful Web applications has become a prominent activity in the software industry during the past decade, recently reaching the GIS field. From a software engineering perspective, the Agile methods and practices emphasise adaptation to these changing circumstances, flexibility, simplicity, usability and user-centered design. Agile development principles have inspired the design of a number of successful open source projects that have proven to be particularly effective in such a mutable and competitive context. For this survey two leading frameworks for Web development, Ruby on Rails and Grails, have been reviewed.

Ruby on Rails is a framework for a rapid Web application development based on the Model-View-Controller paradigm. The project also enforces design principles such as Convention over Configuration and Don't Repeat Yourself (DRY), lowering repetitive operations in the development cycle, driving the developer's attention to the business logic beyond the platform details. The framework models the entire application, from the Web pages down to the underlying relational database, enforcing a rigid structure. The main language used in the framework is the dynamic language Ruby, created by Yukihiro Matsumoto blending parts of Perl, Smalltalk, Eiffel, Ada and Lisp. Ruby's syntax is supposed to directly map natural language and tends to be concise and without frills. Among the packages that compose the framework, ActiveRecord provides an object-relational mapping system for database access, while ActiveResource facilitates the creation of Web services. Ruby on Rails has no native spatial support, although it is possible to map spatial data manually to a spatially-enabled database such as MySQL.

Grails is a Web application framework that closely mirrors the basic design ideas of Ruby on Rails. The main language of Grails is Groovy, a dynamic language for the Java Virtual Machine, designed to be extremely easy to learn for Java developers. Groovy shares a lot of design principles with Ruby but with a Java-like syntax. The architecture of Grails is built on Java-based technologies and claims full compatibility with Java, which is a relevant advantage over Ruby on Rails. Moreover, Grails handles the object-relational mapping with Hibernate. Although Grails has no native spatial support, its compatibility with Java enables the use of several spatial tools, such as the Java Topology Library and Hibernate Spatial.

Object-Relational Mapping

Typically, developers utilise databases to store and maintain persistent objects. However, for this strategy to be successful, consideration of how objects are mapped to the Relational Database Management System (RDBMS) must be considered. Objects are inherently complex, consisting of data combined with state and behaviour and so cannot be stored directly in the database. One of the powers of the object-oriented paradigm is the inheritance model which offers the opportunity for code reuse. This differs from RDBMS whose primary

purpose is to remove redundancy among stored data rather than model real world entities and behaviour.

Within GISs, handling the mapping between the object-oriented paradigm and the relational model is extremely important due to the complex spatial objects which require efficient storage in databases. Several frameworks have been developed to handle this mapping by providing an interface to assist developers, hiding low-level implementation details. In this survey Hibernate and its spatial extension, Hibernate Spatial, are investigated.

Hibernate provides a framework for mapping an object-oriented domain model, in Java, to the traditional relational database paradigm. There are always conceptual and technical difficulties when trying to map objects or classes in Java to relational tables. Hibernate solves these issues by using high-level object handling functions. In addition to handling this, Hibernate can map Java data types to SQL data types which permits Hibernate to include data query and retrieval facilities by generating appropriate SQL queries. This benefits the developer as it removes the task of handling the result set and any object conversions required by the query. Furthermore, one of the main benefits is that it enables applications to function with all the support of an SQL database. In this way, Hibernate permits the development of persistent classes following the object-oriented paradigm, including association, inheritance, polymorphism, composition and collections while providing an interface between Java and any underlying SQL relational database.

Hibernate Spatial. While Hibernate natively supports common data types, spatial entities and geometries are not explicitly handled and require the use of a plug-in called Hibernate Spatial. Hibernate Spatial allows the mapping of these spatial objects to spatial databases which facilitates the use of indices and optimisation. In a similar way to Hibernate, Hibernate Spatial abstracts from the way a specific database supports geographic data, and provides a standardised, cross-database interface to geographic data storage and query functions. Additionally, Hibernate Spatial supports most of the specifications of the OGC Simple Features Interface Standard making it a useful addition for any Java based geospatial Web application using spatial databases. Hibernate Spatial can be integrated with Grails to handle a transparent mapping to PostGIS and other spatial DBMSs.

Spatial DBMSs

A spatial DBMS offers spatial data types in its data model and query language, provides spatial indexing and algorithms for spatial queries. Oracle Spatial, PostGIS and MySQL are some of the popular Spatial DBMS packages in this domain. Advancement and awareness in geospatial technologies has given rise to standardisation and as a result most Spatial DBMSs conform to the Simple Features Interface Standard (SFS). This standard is widely followed by almost all SQL databases and provides a way to access and store geographic data types along with related functions and data operations. In this survey PostGIS and MySQL are examined.

PostGIS is a spatial extension of the PostgreSQL database management system and is a well known open source project for storing and querying geographical entities. PostGIS is widely supported by a large number of proprietary and open source Web mapping servers. It is fully compliant with OGC standards such as SFS for SQL. Apart from the basic spatial support, PostGIS also has topology, data validation, coordinate transformation and programming API functionalities. As the project continues to grow, additional features for raster support, routing, three dimensional surfaces, curves and splines have been proposed.

MySQL is a widely used open source relational database management system, which, since version 4.1, includes support for spatial extensions using an additional geometry column for the storage and analysis of geographic features. These features could be any entity having more than one dimension and associated locational information. MySQL fully implements the SFS and supports R-tree indexing to speed up database operations. It operates on multiple platforms and constitutes a key component of the open source LAMP framework (Linux, Apache, MySQL, PHP/Perl/Python), thus making this a popular open source project.

4.3.2 Online survey design

First-hand knowledge regarding open source projects was obtained from active developers through online questionnaires [311]. Given the nature of open source online communities, such surveys are a particularly suitable way of getting access to the community members. For this purpose, we have designed an anonymous questionnaire with 13 questions, three of which cover the respondent (*Contribution*, *Affiliation* and *Level of expertise*), while ten are related to the characteristics of the project being analysed (*Learning Curve, Stability, Performance, Scalability, Interoperability, Extendibility, Standards, Documentation, Community support, Frequency of updates*). The full questionnaire is reproduced in Appendix A.1.

Each question on the software is measured on a 5-point discrete visual analog scale [296], 1 being the minimum score and 5 the maximum. It is important not to force the responder to give an answer when it is not relevant or interesting to them, so each answer includes a 'no answer' option. In order to maximise the amount of responses, each questionnaire was designed to be completed in less than 90 seconds. Each questionnaire focuses on a specific project, asking the same questions for all the projects, which allows the responses to be compared in a quantitative way. A Web home page, presenting a short description of the survey to the public and linking the 14 questionnaires, was created and published online on April 15, 2010.

The survey was disseminated through several channels with an invitation to take the questionnaires for the technologies that the responder had used directly. The community members were encouraged to take the survey purely to contribute to academic research without any economic incentive. The announcement was posted on the official forums and mailing lists of each project.

The questionnaires were subsequently published on boards for software developers, such as Stackoverflow[13] and on high-activity academic mailing lists such as DBWorld.[14] Responses were received from April 15, 2010 to June 15, 2010.

The opinions collected in this survey represent the subjective and qualitative perception of a number of users in relation to open source online projects. Although online surveys based on self-selection are a cost-effective tool, they have some drawbacks [311]. For example, the demographics are self-assessed, so a responder could easily provide false information about their affiliation and contribution to the project. However, given the context, this is unlikely to happen because there would no benefit for the responder. The responders intrinsically form a non-probability sample whose representative accuracy is difficult to measure. The object of this survey is a set of open source projects and not their online communities, therefore the collected feedback still offers valuable information about the software characteristics. Our questionnaire is based on discrete visual analog scales, which can be interpreted in different ways by different responders. In order to minimise this problem, the scales were labelled as clearly as possible, avoiding jargon and uncommon words, which could be problematic for non-native English speakers.

The responses obtained from the online survey are described in Section 4.3.3 mainly through average values. When using traditional Likert scales to obtain feedback from respondents, the average value does not provide a meaningful description of the results because the middle answer is usually labelled 'neither agree nor disagree.' On the contrary, the middle answer of the visual analog scales used in our online questionnaire represents the actual midpoint of the scale (3 on scale from 1 to 5), with a separate 'no answer' option. Therefore, in this questionnaire average values will show overall trends in the results.

In the literature, several categories of bias have been identified in online surveys. For example, a 'central tendency bias' occurs when respondents avoid extreme responses [69]. Similarly, a positive ideological bias has been identified within open source communities [118]. However these aspects were beyond the scope of our research. Further statistical analysis could help understand them, for instance by taking into account the correlation between the responder's contribution and the scores they express (e.g. an attempt at publicising a project the responder has worked on). Although these drawbacks exist, the questionnaire results provide valuable indicators about the project characteristics based on users' perceptions.

4.3.3 Online survey results

This section reports the analysis performed on the survey results, including the demographics and the overall scores. Section 4.3.4 will focus on a visual representation of the results for each project.

[13] `http://stackoverflow.com`
[14] `http://www.cs.wisc.edu/dbworld`

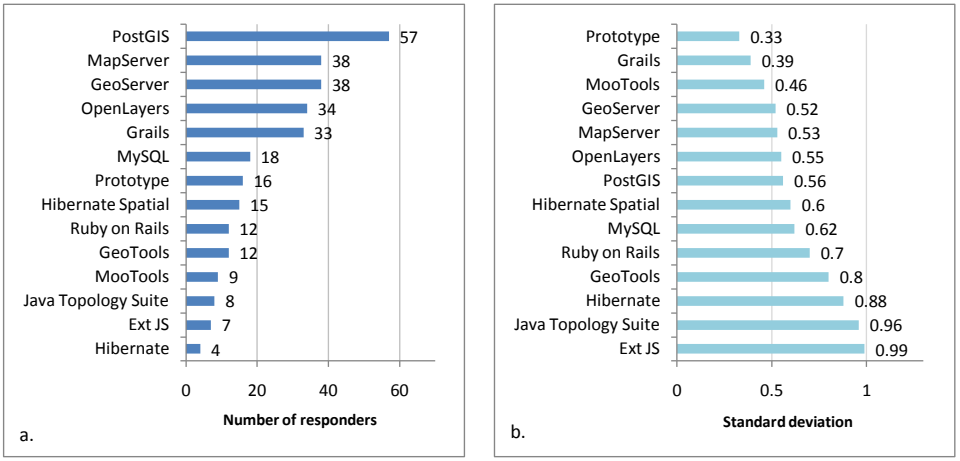

Figure 4.1: (a) Number of responders [total: 301] (b) Standard deviation per project

In total, the 14 questionnaires obtained 301 responses. Due to the nature of online surveys, it is not possible to generalise the results from the sample to the entire relevant community. Online surveys based on self-selection produce non-probability samples [311]. Although these results might not express valid general knowledge on those online communities, which are not the subject of our survey, they can be used to reinforce a qualitative description of the projects being examined. In this analysis the responses are treated as independent. Figure 4.1a shows the distribution of responses over the different projects.

There seems to be no correlation between the community size and the number of responses. In certain cases, few members of large communities have responded (e.g. Ext JS), while small communities showed a higher interest in the survey, perhaps in an attempt to increase the visibility of their project (e.g. Hibernate Spatial). Given the uneven distribution of number of responders, which varies from 57 to 4, it is beneficial to display the standard deviation (figure 4.1b). The standard deviation and the correspondent standard error in this context represent the measure of precision of a specific project. Intuitively, the precision of a score calculated over a high number of responses is likely to be higher than that over a very small sample.

In terms of demographics, we consider three self-assessed characteristics of the responders: affiliation, contribution and level of expertise. Of the 301 responders over all 14 projects, 54 have said they are contributors (18%), while 244 are non-contributors (81%) and 3 did not reply (1%). Similarly, 62 consider themselves as affiliated to the project (20%), 231 non-affiliated (77%) and 8 did not answer (3%). The levels of expertise have the distribution displayed in Figure 4.3. It is possible to observe that most responders evaluated their expertise in relation to the project as 3/5 (36%) or as 4/5 (29%).

Figure 4.2a represents the scores of all the projects sorted from the highest to lowest value. The score of a project is the average of the average of each response. Grails, PostGIS and MooTools have obtained the highest score, while Hibernate Spatial, Hibernate and GeoTools obtained the lowest.

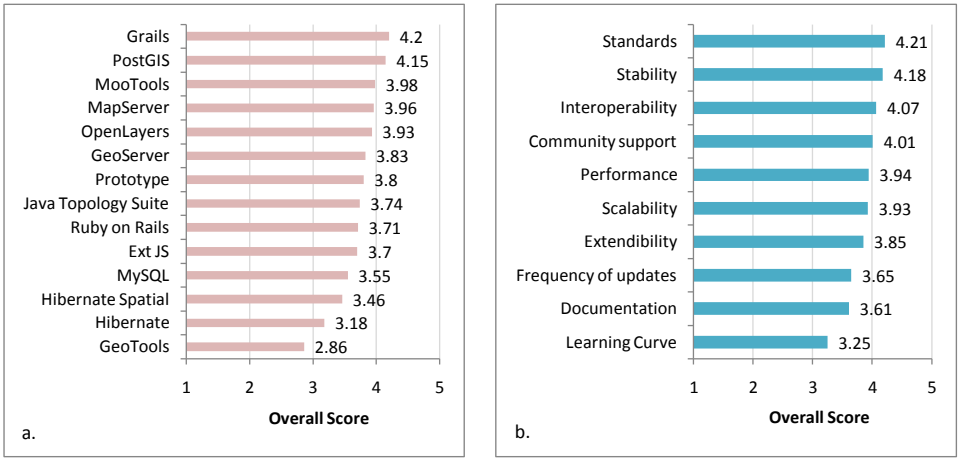

Figure 4.2: (a) Overall scores for projects (b) Overall scores for characteristics

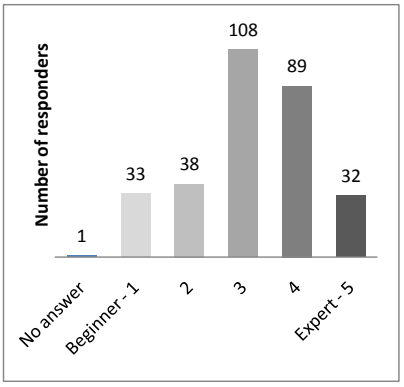

Figure 4.3: Level of expertise distribution [total: 301]

The score of a characteristic is the average of the responses across all the projects. Figure 4.2b summarises these scores sorted from the highest to lowest value. Support for standards, stability and interoperability have been ranked the highest. The responders have deemed the frequency of updates, the quality of documentation and the learning curve to be the least satisfactory characteristics. This result confirms the low priority that usability has for open source developers, who generally devote more resources to other goals, such as software quality [10, 173].

4.3.4 Technology comparison

In the survey, feedback was obtained on 14 projects using the ten characteristics outlined above. The average score of each characteristic was calculated for each individual project, resulting in 140 scores. In order to visualise these values, which represent non-comparable features on the same scale, radar charts are particularly suitable. Therefore a project is represented by a polygon whose vertices indicate the average score for the ten characteristics. The origin of the chart, in which all the axes converge, corresponds to the minimum possible value of the score (1), while the outer end points of the axis show the

maximum score (5). Therefore, the shape of the polygon reflects the scores: a project that obtained balanced scores tends to be a regular decagon, whilst uneven scores generate spikes and valleys in the polygon. Similarly, a small polygon represents a set of lower scores than a large polygon.

Figure 4.4 offers a visual representation of the opinions received from the responders in relation to the client-side technologies (OpenLayers, Ext JS, Prototype and MooTools). When the Web GUI libraries are examined, it can be seen visually that they generally follow a similar pattern. The same can be said about the AJAX libraries (Prototype and MooTools). The only noticeable variation is a spike in the frequency of updates, in which Ext JS performed better. The results for the Web mapping servers and spatial libraries (GeoServer, MapServer, GeoTools and Java Topology Suite) are outlined in Figure 4.5. For the majority of characteristics MapServer obtained more positive feedback than GeoServer. The difference between the spatial libraries is more pronounced, as GeoTools received lower scores than the Java Topology Suite.

Figure 4.6 provides the radar representation of the results obtained in relation to the server-side technologies (Ruby on Rails, Grails, Hibernate and Hibernate Spatial). It is evident that Grails received noticeably higher scores than Ruby on Rails, except in relation to stability. Additionally, a valley is present in the community support for Hibernate. The results related to the spatial DBMSs are provided in Figure 4.7. While both projects are seen as having a similar learning curve, the opinions about PostGIS are noticeably more positive than those of MySQL.

A visual representation of the scores for each project allows the reader to compare their characteristics for use in different application domains. As the use of open source geospatial technologies increases, these results are applicable to a wide range of domains where interactive maps are required. Many traditional online services are now being augmented with spatial components. For Web developers this poses a challenge, as they need to determine the appropriate technologies to adopt. Through the survey presented in this article, the core technologies are evaluated by 301 real users. The feedback collected offers first-hand information about the characteristics of the projects.

The results can be used by developers in any field who wish to include spatial support in their projects. For example, when evaluating alternative spatial DBMSs, a developer can take into account the results of our survey, which indicate that developers are more satisfied with PostGIS than MySQL. Similar conclusions can be drawn for all 14 technologies analysed. The results of this survey can also be considered by the project contributors to address issues in order to improve the overall developers' experience. The scores obtained from the responders confirmed the choice of technologies for the implementation of the prototype of our Web architecture for open GIServices [200]. The Web mapping server MapServer, the Web application framework Grails and the spatial DBMS PostGIS, which form the core components of our architecture, received a higher overall score from the responders than similar projects (GeoServer, Ruby on Rails and MySQL respectively).

This work can be extended to include staff from the companies and or-

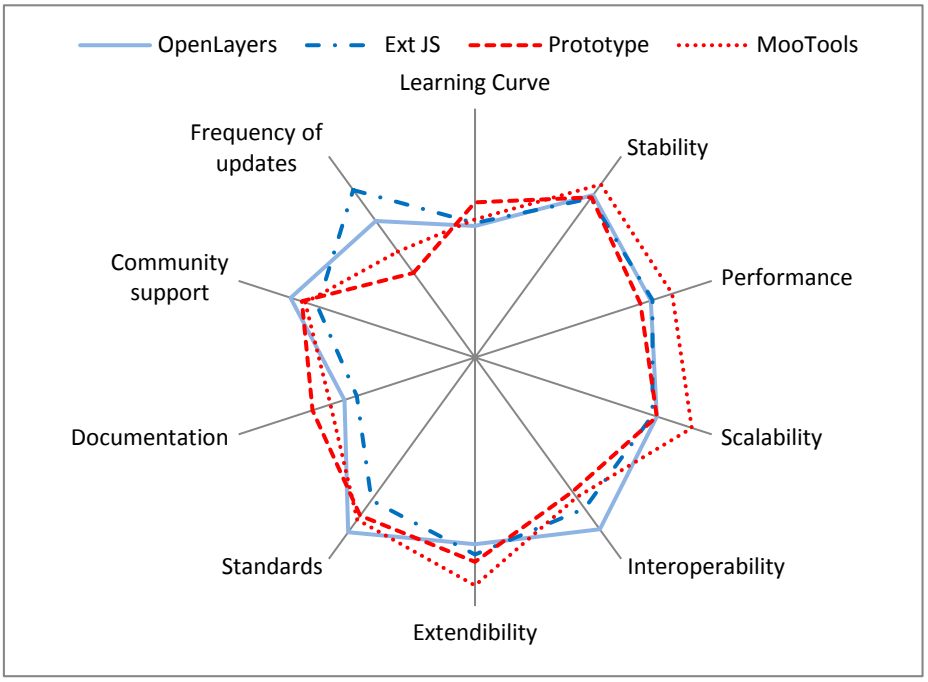

Figure 4.4: Web GUI and AJAX libraries

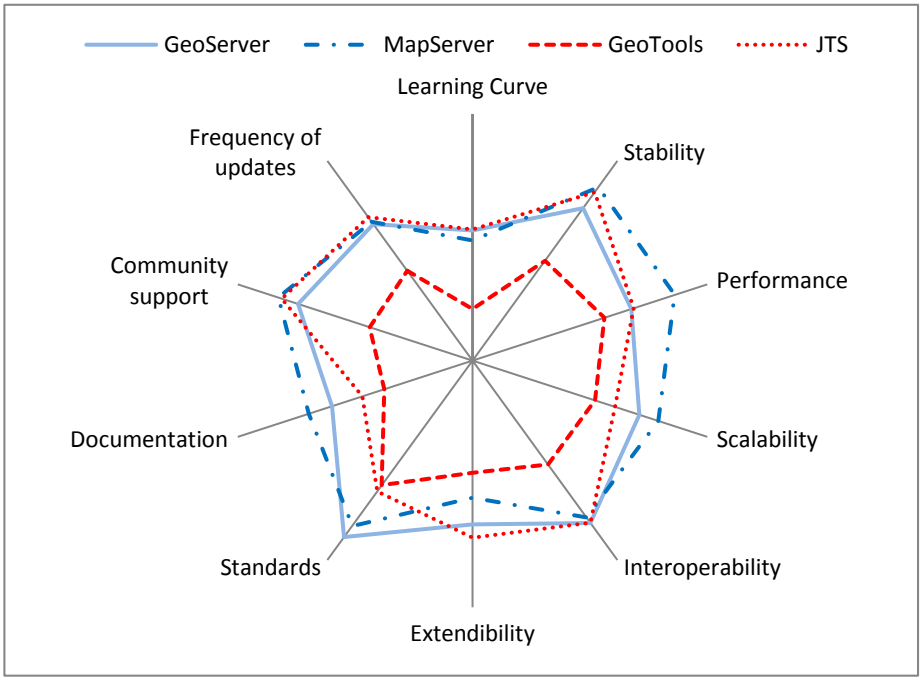

Figure 4.5: Web mapping servers and Spatial libraries

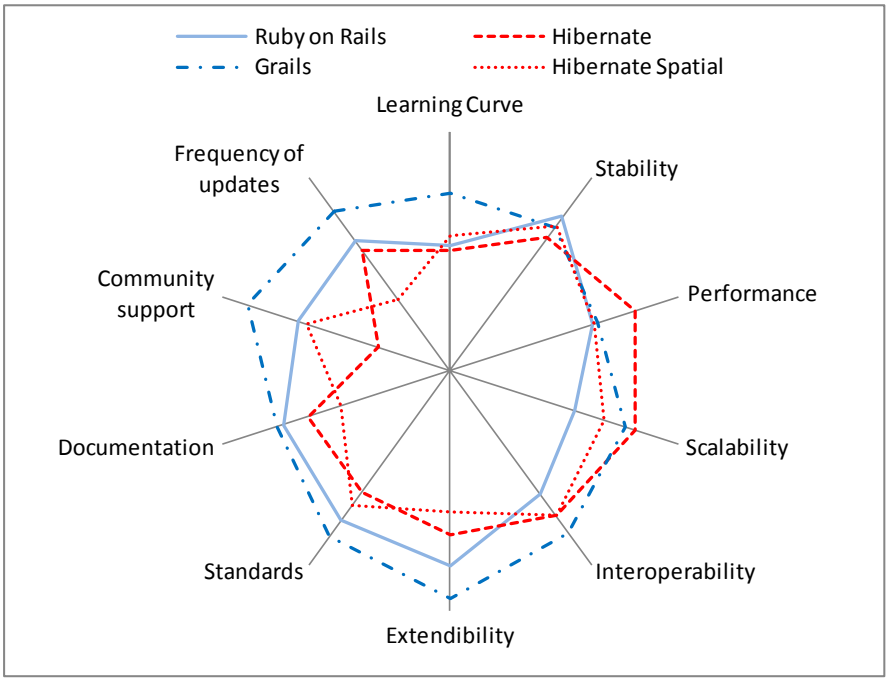

Figure 4.6: Web application frameworks and Object-Relational mapping

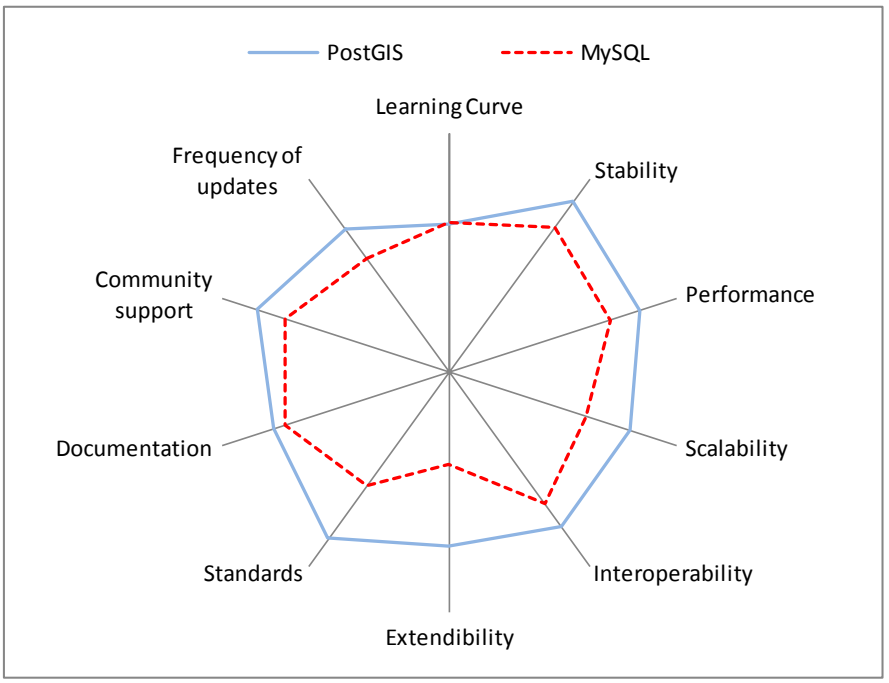

Figure 4.7: Spatial Database Management Systems

ganisations, thereby providing better control over the responders' demographics. The results obtained through the discrete visual analog scales can also be analysed using alternative statistical techniques. Thus, the correlation between expertise level and the responses could be investigated, as well as the variation between contributors and non-contributors. Furthermore, the comparison between this study and surveys based on other approaches could reveal interesting trends. The developer's perception contrasted with findings from other sources might also assist in understanding further the relationship between human users and software tools. Based on the opinions expressed by 301 users in this survey of 14 open source Web mapping projects, the following conclusions can be drawn:

- In the users' view, free and open-source software (FOSS) for Web mapping tends to provide good support for existing standards, stability, interoperability with other systems, offering community support.
- Intermediate satisfaction is expressed in relation to performance, scalability, and extendibility of the projects. By contrast, users expressed lower satisfaction for user friendliness, quality of the documentation, and the frequency of updates.
- Grails, PostGIS, MooTools, MapServer, and OpenLayers obtained the best overall user scores. Open source tools provide components that can be used to construct Web architectures for Web mapping, and other geographic applications (see Section 4.4 for an example).

Web developers can take these indications into account when choosing suitable tools for Web mapping applications. Similarly, project owners and contributors may address the weaknesses that the users have identified in these projects, such as low user friendliness and poor documentation. These results informed the development of our Web platform for map personalisation and visualisation, presented in the next section.

4.4 Web platform for map personalisation and visualisation

In order to explore map personalisation from a semantic perspective, we have implemented an open source Web platform for map personalisation and visualisation. This platform has the purpose of providing a testbed for spatial personalisation and recommendation, in a consistent and standardised way, extending the RecoMap prototype to a Web-based architecture (see Section 3.2.1). This work was conducted in the framework of Strategic Research in Advanced Geotechnologies (StratAG), and was published in [200].

The system proposed is an open web platform for spatial personalisation and visualisation. Based on a client-server architecture, the system delegates the task of computing complex algorithms, generating visual output and dealing with costly operations to the server. The clients, on the other hand, send

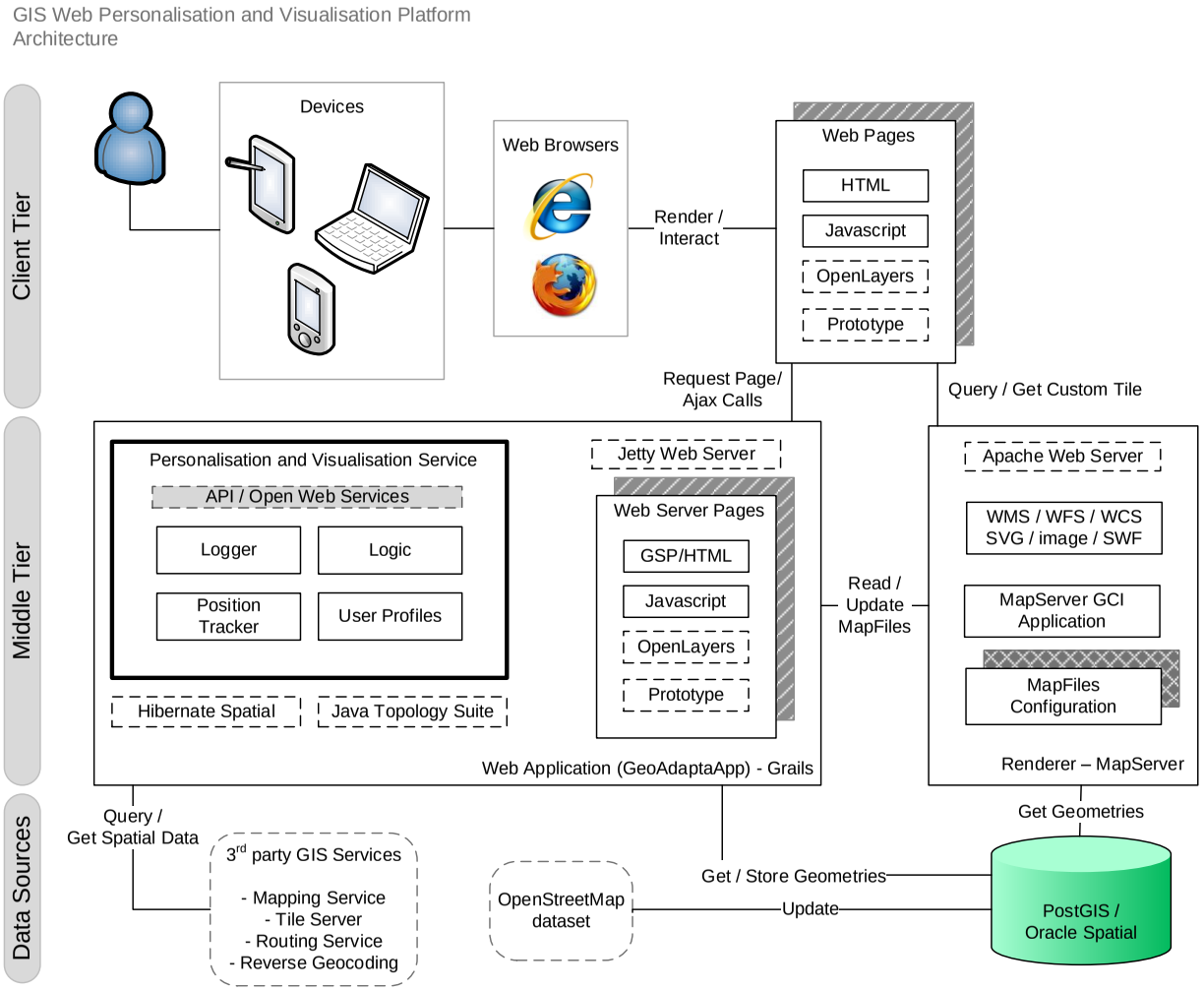

Figure 4.8: Web platform for map personalisation and visualisation: system architecture

requests to the server and render the output for the user. Within this domain interoperability is crucial. Firstly, input and output spatial data have to be defined in well-known formats. Secondly, the web-based approach minimises the coding overhead necessary to port the application to different mobile and desktop platforms. Last but not least, Open Web Services and an API expose the system functionalities on the Internet and allow external applications to interact with the system.

The architecture of the system proposed is structured in three tiers: client tier, middle tier and data sources. This approach emphasises the independence of the various components of the system, that can be deployed and combined in different contexts. Figure 4.8 illustrates the interaction between tiers, outlining the main components of the system and their logical position. The following sections describe these tiers in detail.

4.4.1 Client Tier

The user interaction with the system takes place in this tier. The user visualises Web pages containing spatial information on their device through a common web browser. These web pages display interactive maps and monitor certain actions performed by the user, such as mouse clicks, zoom, etc. Such actions get sent to the Personalisation and Visualisation Service. In a typical scenario a web page displays a dynamic map served by the Map Renderer and other information from the Personalisation and Visualisation Service (both situated in the Middle Tier).

4.4.2 Middle Tier

This tier contains the core services and functionalities of the system. A web application hosts the Web Server Pages which constitute the access point to the system for the users and the Personalisation and Visualisation Service, in which several web services are deployed and exposed on the Internet.

This Personalisation and Visualisation Service tracks the user sessions, handles the user profiles, logs all of the relevant actions performed by the clients and keeps track of the user location when available. The personalisation and visualisation algorithms are also implemented within this service to take advantage of the server-side computational power and full access to the spatial datasets and recorded interaction. The Personalisation and Visualisation Service can be accessed by external applications through an API defining functions with parameters and the format of the returned values. The Web Server Pages define the dynamic contents that are served to the clients. A Web Server Page can aggregate spatial and non-spatial data from heterogeneous sources, tailored for a specific user. Such a page gets rendered and delivered to the client. The navigation logic between these pages is defined and controlled by the Web Application.

At the same architectural level a Map Renderer is deployed. This component is able to load spatial data and to render them according to a dynamic

map file configuration, which defines what data have to be rendered and with which visual style. The map file configurations are generated dynamically by the Personalisation and Visualisation Service according to the profile of the user receiving the spatial data. The Map Renderer hosts several map servers, which are spatial web services specifically designed to provide the clients with spatial information in standard and widely-adopted formats. Both the Personalisation and Visualisation Service and the Map Renderer interact with the Data Sources to obtain and store the spatial information needed to execute their tasks.

4.4.3 Data Sources

The main function of this tier is to provide the other tiers with spatial information. Its main component consists of a spatial DBMS, which stores spatial datasets (e.g. vector maps) and data related to the user profiles. The spatial DBMS is used by the Personalisation and Visualisation Service to run the personalisation logic, whilst the Map Renderer accesses it to perform the visualisation. The relevant spatial datasets can be imported from an external data provider (e.g. OpenStreetMap), and stored in the spatial DBMS and updated on a regular basis.

Another apparent phenomenon on the Internet is the growing availability of commercial GIServices accessible for free (e.g. Virtual Earth, Google Earth, Google Maps and CloudMade). These services rely on high-end commercial infrastructures to provide access to various datasets through Web APIs, mostly for routing, geocoding and reverse geocoding. The Web architecture described in this section lets the Web Application request data from such web services with low development costs.

4.4.4 Personalisation algorithm

The personalisation algorithm described in [20] has been implemented in the Personalisation and Visualisation Service. The user sees the world as a set of 'items.' An item is a point of interest representing a geographical entity (either a point or a polygon) which the user might have an interest in. Certain types of item (such as shops, cinemas, etc) can be associated with relevant resources on the Web (such as web sites or web services). These items are displayed on an interactive digital map, rendered by the Map Renderer.

The Personalisation and Visualisation Service assigns an interest score to each item located within the current *interest radius*. The interest radius is inversely proportional to the current user speed. The interest score α for the item i is calculated with Equation 4.1 taking into account both historical interactions (*interaction*) and current user routing distance from the item (*proximity*)

$$\alpha_i = P_i P_R + I_i I_R P_R + I_R = 1, \quad P_i \in [0,1], \quad I_i \in [0,1] \tag{4.1}$$

P_R and I_R are respectively the *proximity ratio* and the *interaction ratio*,

meaning the weight the system attributes to each indicator. Those ratios are dynamic and change over time: the more the user interacts with items within the interest radius, the more the scales will be tipped toward P_R to emphasise proximity, and vice versa. Furthermore, the proximity score P_i and the interaction score I_i are normalised between the maximum and minimum score among the items within the proximity score. Given the volatile nature of user interests [312], a time decay function based upon the days elapsed since the last interaction with the item i is also applied on $|I_i|$ and $|P_i|$.

When a certain condition occurs in the user context and profile (e.g. user moves to a new area), the Personalisation and Visualisation Service triggers an adaptive action. For example, when the user either starts to explore a new area or alters the user profile by interacting with the map, the client requests new recommended items. The items having the highest interest score at a given time for a given location are sent to the map.

4.4.5 Semantic service

A semantic component was integrated into the architecture. The additional semantic support consists of the technique to enrich OSM data, described in Section 3.2.2. The semantic module aims at extending semantically-poor OSM vector data with semantically-rich datasets, such as DBpedia, exploiting the mapping provided by LinkedGeoData. Figure 4.9 shows the integration with the semantic module, including the open source technologies empowering each layer.

Our Web platform for map personalisation and visualisation is a tiered architecture, offering a testbed for map personalisation techniques. The client layer offers a Web mapping GUI, constantly monitoring user interaction (i.e. mouse clicks and movements for traditional clients, taps for touch-based devices). The middle layer hosts the core functionality and personalisation logic, storing the user profile. This layer draws on several VGI data sources, to gather and display relevant spatial information to the user, tailoring it for their specific interests. While developing this Web platform for map personalisation and visualisation, we have identified the semantic gap in OSM as a critical issue to be tackled, which then became the main focus of this thesis. As part of this work on the shared Web platform, we carried out a survey of open source tools for Web mapping, reported in the next section.

4.5 Summary

In this chapter, we have focused on two aspects of our contribution. First, we presented the OSM Wiki Crawler, a tool to extract a novel resource, the OSM Semantic Network (Section 4.2). This Web crawler scans the OSM Wiki website, where OpenStreetMap (OSM) contributors define and discuss their conceptualisation of geographic objects, which are then utilised in the OSM vector data. Heuristics are utilised to extract a Resource Description Frame-

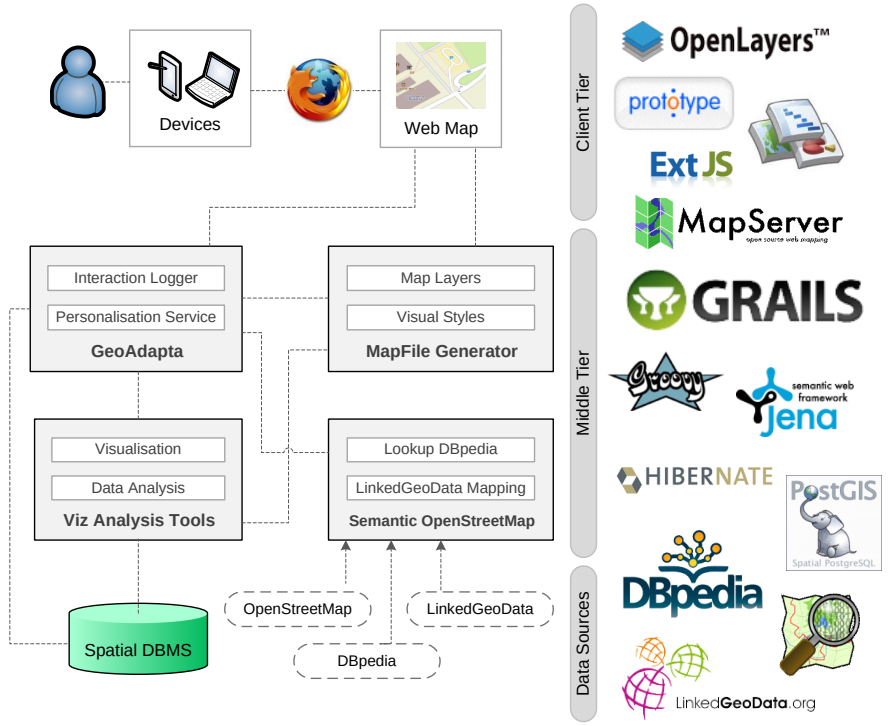

Figure 4.9: Web platform for map personalisation and visualisation: semantic integration

work (RDF) graph of relevant knowledge about the thousands of geographic concepts that provide the core of OSM semantics.

Second, we described our preliminary work on map personalisation and implicit feedback, conducted in collaboration with the National University of Ireland, Maynooth, in the framework of the Strategic Research in Advanced Geotechnologies (StratAG) cluster.[15] This work was crucial in identifying the research problem that we tackle in this thesis. In the context of this collaboration, we designed a Web platform for map personalisation and visualisation, relying on implicit feedback to provide a range of personalisation services. As part of the design process, we investigated the free and open-source software (FOSS) ecosystem, in the rapidly changing landscape of geospatial Web technologies, with the purpose of selecting the optimal set of technologies, as well as providing a useful indication to other potential developers.

Thus, we carried out a user study, involving relevant communities of open source developers (Section 4.3). This study included a variety of open source geo-Web technologies, such as Web application frameworks, spatial libraries, and Web GUI toolkits. The responders expressed higher satisfaction for the stability and degree of standardisation of such tools. User-friendliness and documentation, by contrast, were largely seen as less satisfactory. The structure of the resulting Web platform for map personalisation and visualisation is described in detail (Section 4.4).

[15] `http://www.nuim.ie`, `http://www.stratag.ie`

As discussed in previous chapters, the OSM Semantic Network is a semantic support tool for OSM, which can be used to compute semantic similarity of concepts. In Chapter 3 we described our approaches to this useful semantic task. An extensive evaluation has to be performed, with the purpose of validating the different techniques considered. The next chapter reports several preliminary experiments that were conducted on the topic of semantic similarity, collecting a body of empirical evidence that highlights the strengths and weaknesses of each technique.

CHAPTER

FIVE

PRELIMINARY EVALUATION

5.1 Overview

The main outcome of Volunteered Geographic Information (VGI) is a growing corpus of open, crowdsourced geographic data, collected by a diverse group of user/producers. In open, dynamic semantic models, enabled by Web 2.0 technologies, the meaning of concepts is negotiated among contributors, resulting in complex, ambiguous semantic content of varying quality. In this thesis we concentrate on the semantic model of OpenStreetMap (OSM), the leading grassroots mapping project. To fill the semantic gap on the OSM concepts, we have developed the OSM Semantic Network, and a hybrid technique for semantic similarity for geographic concepts, the Network-Lexical Similarity Measure (NetLexSiM), both described in Chapter 3. The main purpose of this chapter is a preliminary evaluation of the two aspects of NetLexSiM (network and lexical), adopting an existing similarity gold standard, the MDSM evaluation dataset.

Empirical evidence is not equivalent to formal proofs, and is always collected in a uncertain context, and can be affected by several biases and methodological issues [306]. Computer science research has historically favoured an engineering approach (the design and development of applications, demonstrations, etc.), and mathematical ones, i.e. formal proofs, showing a certain degree of scepticism towards experimentation with human subjects. However, as many have pointed out [287, 302], empirical experimentation is crucial not only to identify design issues and to validate hypotheses, but also to suggest new theoretical perspectives, in a virtuous interplay. For this reason, we devoted efforts to the empirical evaluation to validate and discuss the contributions proposed in this thesis.

This chapter is structured as follows. First, we report a preliminary evaluation of our technique to enrich the semantics of OSM, using open knowledge bases such as DBpedia and LinkedGeoData (Section 5.2). The remainder of the chapter is then devoted to the pilot evaluation of NetLexSiM, our technique to compute the semantic similarity of concepts contained in the OSM Semantic Network, based on network and lexical similarity. The purpose of this pilot evaluation consists of analysing the performance of the technique with a combination of parameters, pointing out the most promising techniques that can be used to compute the semantic similarity of OSM concepts. This evaluation

is grounded on the assessment of cognitive plausibility, i.e. the correlation between human and machine-generated similarity judgements (Section 5.3).

To conduct this pilot evaluation on semantic similarity, we rely on an existing gold standard, the MDSM evaluation dataset (Section 5.4). The pilot evaluation of network-based algorithms is performed, reaching the highest cognitive plausibility for the algorithm SimRank [136] (Section 5.5). The same gold standard is also used for the evaluation of our approach to lexical similarity, based on the crowdsourced definitions of geographic concepts, which obtains the highest plausibility for a combination of the text similarity measure proposed by Corley and Mihalcea [54], and the WordNet-based measure by Jiang and Conrath [137] (Section 5.6).

Conducting the pilot evaluation on the lexical component of NetLexSiM, we developed the analogy of the 'similarity jury,' which compares similarity measures to members of a panel of domain experts, judging the similarity of a pair of concepts (Section 5.7). In a domain where no gold standard is available, the jury provides a reliable approach to measure semantic similarity of concepts.

5.2 Evaluation of OSM semantic enrichment

The OpenStreetMap (OSM) vector dataset contains a vast amount of spatial information, based on a shallow semantic model, describing geographic entities via tags. To extend the OSM semantic content, we have developed a semantic service that integrates OSM data with the non-spatial ontology DBpedia (see Section 3.2.2). This section describes a preliminary evaluation that we have performed to assess the performance of the system, highlighting a semantic gap in the representation of OSM concepts, which ultimately led to the development of the OSM Semantic Network.

In order to analyse the system behaviour on large geographical areas, we have built an interface to perform sampling experiments. Through the page represented in Figure 5.1, the system can analyse a specific region of the map. The user can draw a bounding box on the map and choose several parameters of the sampling to be performed. The controls in the *ontologies* section of the page allows the user to select which ontologies have to be included in the experiment. In order to contrast the results, it is possible to enable or disable the LinkedGeoData mapping and the DBpedia lookup service.

The points to be analysed can be selected in two modes: *random* and *grid*. The former consists of selecting a number of random locations within the bounding box, while the latter distributes points on the bounding box at a regular distance (specified in the parameter *grid distance*, expressed in degrees). The map scale can either be set to a given value for all the points, or can be selected randomly for each point. In this way it is also possible to observe what the impact of the scale is on the results.

The experiment can be executed with the button *launch experiment*. The system then executes a query on each point with the selected parameters, and

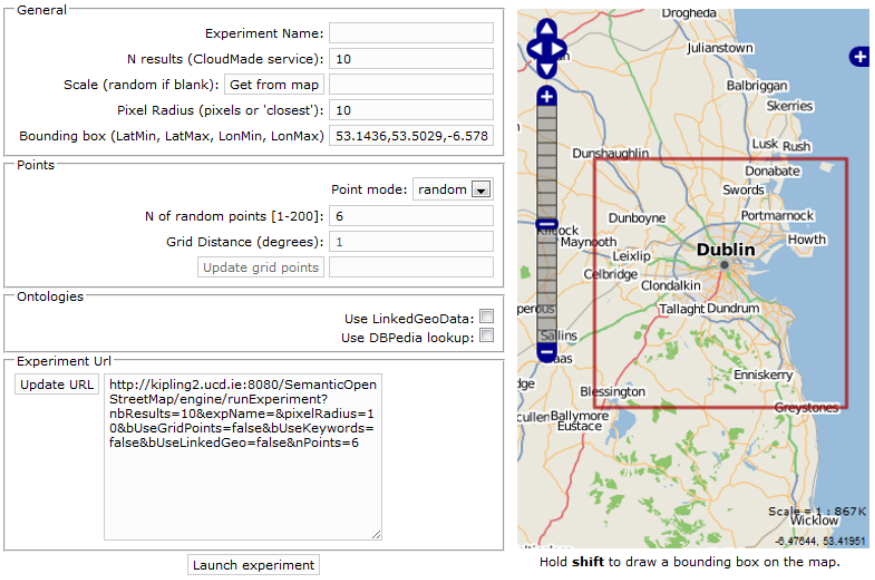

Figure 5.1: GUI for sampling a bounding box by running multiple queries

Figure 5.2: Grid sampling a bounding box surrounding the Dublin area

all the results are merged. For the moment the complexity of the procedure is linear, so the time grows linearly with the amount of selected points. When completed, the experiment results are stored in the XML file containing all the spatial queries and the relevant ontological entities. The input and output of each query are included in the file, in order to be easily machine-processed and analysed. At the end of the procedure, the URI of the XML document is returned to the user, and the file can be used for any other analysis.

Figure 5.2 shows an example of grid sampling of a large bounding on the Irish East coast with 12 points. When executing a small experiment with scale set to 14,000, radius equal to 10 pixels (equivalent to 50 meters at this scale), the results contain 13 DBpedia nodes, 43 categories and 27 ontological terms. One of the retrieved resources is the Irish village Roundwood, stored in the XML code below. The resource contains an OSM structure (*osm*) and a DBpedia node (*dbpedia*), with a geo-location, one category and three ontological terms:

```
<resource index='0'>
  <osm id='52270446' name='Roundwood'>
     <location lon='-6.22481' lat='53.06347' />
  </osm>
  <dbpedia>
     <node>Roundwood</node>
     <location lon='-6.2333' lat='53.06' />
     <subjects>
      <category>Towns_and_villages_in_County_Wicklow</category>
     </subjects>
     <ontology>
       <term>Municipality</term>
       <term>Place</term>
       <term>PopulatedPlace</term>
     </ontology>
  </dbpedia>
</resource>
```

This code can be easily processed and the relevant URI[1] can be accessed. Thus, links between the spatial vector data displayed to the user and the ontological and semantic richness of DBpedia have been enabled.

During the development of the system we carried out a preliminary evaluation with postgraduate students of the School of Computer Science and Informatics, University College Dublin. In order to engineer the heuristics, and establish a suitable value for the thresholds ϵ (the maximum node distance) and σ (the tag matching ratio), we defined three alternative versions of the system, combining the phases in different ways:

- *Version 1*: OpenStreetMap + LinkedGeoData mapping
- *Version 2*: OpenStreetMap + Our heuristics
- *Version 3*: OpenStreetMap + LinkedGeoData mapping + Our heuristics

Versions 1 and 2 are using only certain parts of the system, while Version 3 exploits its full functionality, similar to the approach used by Mirizzi et al. [208]. The following parameters have been defined as constant: ϵ (maximum node distance) = 50km, σ (tag matching ratio) = 0.5, *maximum results* (CloudMade service) = 10, *radius* = 20 pixels, *scale* = random, *bounding box* = Dublin area.

The users were presented with a set of six random spatial queries and the corresponding view on the CloudMade map scaled and centred to the query location. The queries had been performed with different versions of the algorithm (two for each version), hiding this information from the user. The users were then asked to rank the semantic relevance of each node (either DBpedia entity or ontological term) on a Likert scale from 1 to 5 (from 'strongly uncorrelated' to 'strongly correlated' with the visible map).

[1] `http://dbpedia.org/resource/Roundwood`

During this preliminary evaluation, several issues were identified. Little difference separated the performance of Versions 2 and 3. The LinkedGeoData mapping, although very accurate when present, did not significantly improve the results, which were mostly returned by our heuristics based on DBpedia lookup. In some cases, such as for the term *Smithfield*,[2] which identifies numerous places in former British colonies, the heuristics succeeded and selected the correct neighborhood in Dublin, based on the geo-location. However, certain mismatches still occurred. When the OpenStreetMap feature contains a very frequent term in a certain geographical area and does not have other semantic content, a mismatch is likely to occur. For example, this typically happens with common surnames (e.g. *Smith*) that are highly frequent in the datasets. Overall, on 47 retrieved nodes, 4 were considered as irrelevant (8.5%).

Although this preliminary evaluation has confirmed that the system works correctly in most cases, the sample is too small to draw general conclusions about it. The objective of this evaluation was mainly to identify design flaws that needed to be addressed. In order to draw general conclusions about the system performance, an extensive evaluation is needed. In particular, more independent variables need to be defined, such as the thresholds, the radius and the maximum number of results, and a bigger sample of queries must be taken into account. This way, several aspects of the system behaviour can be better assessed and quantified.

This preliminary work on the semantic enrichment of OSM (see Section 3.2.2) was crucial to identify the lack of semantic support focused on the OSM concepts. This semantic gap then became then the focus of our work, leading to the development of the OSM Semantic Network, extracted from the OSM Wiki website. The remainder of this chapter is devoted to the pilot evaluation of the Network-Lexical Similarity Measure (NetLexSiM), our semantic similarity measure for OSM geographic concepts. The next section will discuss the strategy we adopt to evaluate NetLexSiM.

5.3 Cognitive plausibility

This section discusses the approach we adopt to evaluate our approach to computing the semantic similarity of OSM geographic concepts, NetLexSiM. In general, measures of geo-semantic similarity and relatedness assign scores to pairs of geographic concepts based on specific criteria, enabling a number of computational applications [134]. Two complementary approaches to evaluating similarity measures have been widely utilised: *cognitive plausibility* and *task-based evaluation*.

The approach based on cognitive plausibility aims at quantifying the effectiveness of a similarity measure through psychological tests, directly comparing human similarity judgements on set of pairs with machine-generated scores. Janowicz et al. [133] define the cognitive plausibility as "how well the rankings computed by the similarity theory match human similarity judg-

[2] `http://dbpedia.org/page/Smithfield`

ments" (p. 115). The psychological evaluation has a long tradition, dating back to the evaluation adopted by Rubenstein and Goodenough [252]. This approach to evaluation is inscribed in the tradition of artificial intelligence that Russell and Norvig [253] define as the ability of 'acting humanly,' i.e. mimicking human behaviour regardless of the actual operations performed by humans.[3]

Task-based evaluation, by contrast, applies the similarity measure to a specific task, for example in the area of natural language processing. The performance of the similarity measure is inferred from how satisfactorily the task is carried out, using appropriate information retrieval (IR) metrics such as precision and recall. A typical task in this area is that of paraphrase detection, aiming at identifying sentences that express the similar concepts with different terms [54, 75]. Ontology alignment tasks, such as those proposed by the Ontology Alignment Evaluation Initiative,[4] can also be used to indirectly assess the effectiveness of semantic similarity measures [68].

The intrinsic high subjectivity of similarity rankings makes the collection and validation of gold standards complex and challenging (see Section 6.2). Although task-based evaluations might appear more 'objective', they are equally affected by subjectivity: ultimately, similarity-based tasks are generated, executed, interpreted, and validated by human subjects. Hence, the reliability of a similarity evaluation should be grounded in stability over time, consistency across different datasets, and reproducibility of psychological results. If performed soundly, both evaluation approaches should show convergent, cross-validating results: a strong correlation is expected between the cognitive plausibility of a measure and its performance in similarity-based tasks. Ideally, both approaches should be adopted to assess the merits and flaws of a similarity measure, starting from cognitive plausibility, and then applying it to a range of more practical tasks, narrowing the observation down to specific contexts. In our work on semantic similarity for OSM geographic concepts, we assess the cognitive plausibility of the proposed approaches, leaving task-based evaluations as future work.

As an indicator of cognitive plausibility, a measure of correlation is necessary. Therefore, the choice of an appropriate statistical test of correlation is crucial, as it has a strong impact on the results. Four statistical tests are able to assess the correlation between two variables, i.e. the machine generated and the human generated similarity scores: Pearson's r, Spearman's ρ, Kendall's τ, and Goodman and Kruskal's *gamma* [276, 148, 99]. Early studies used Pearson's r between the similarity scores and the gold standard [249, 245, 1]. Given that similarity functions cannot be assumed to be linear, Pearson's r was progressively discarded.

Most studies on semantic similarity report Spearman's ρ or Kendall's

[3]The term 'cognitive plausibility' is often used in the scientific literature, and we will adopt it too. However, it suggests a misleading emphasis on human cognition, which is normally beyond the scope of computable approaches to semantic similarity. As these evaluations only compare the visible behaviour of human subjects and computable similarity measures, the term 'behavioural plausibility' could be a more accurate alternative.

[4]`http://oaei.ontologymatching.org`

τ, popular non-parametric measures that certainly seem more suitable than Pearson's r [276, 232, 314]. The two correlation coefficients share the same assumptions about the data, but differ considerably in their approach: while ρ is simply Pearson's r computed on the score rankings, τ represents the difference between the probability of the variables being in the same order and the probability of this not being so. This difference notwithstanding, ρ and τ are *equivalent*, i.e. they are within a constant multiple of each other [72]. For this reason, there are no strong arguments to prefer one or the other. Moreover, ρ is generally higher than τ [269]. This makes ρ more attractive to researchers who are trying to maximise such correlations, and can account for the fact that it is used more frequently than τ.

The ranking comparison needs extra caution in the case of partial rankings, that is, rankings having *ties*. For example, the WordNet-based lexical measures by Rada et al. [239] and Hirst and St-Onge [124] tend to generate a large number of tied pairs. The original versions of ρ and τ do not take ties into account, but corrected versions have been devised [57]. When the number of tied rankings is very high, Goodman and Kruskal's *gamma* is the most appropriate measure [99]. Several other measures for partial rankings have been devised by Fagin et al. [71, 72].

With the exception of Pearson's r, all the aforementioned coefficients could reasonably be adopted to compute the correlation between human and machine-generated rankings. However, in order to make our results comparable with the body of studies on semantic similarity and relatedness, we adopt a tie-corrected Spearman's ρ, as utilised by Rodríguez and Egenhofer [250]. In our experiments, tied rankings are corrected through the mean ranking. For example, scores $.22, .31, .31, .47$ are ranked as $1, 2.5, 2.5, 4$. For all the experiments described in this chapter, we also computed Kendall's τ, always obtaining a very strong correlation with the corresponding Spearman's ρ (both ρ and $r > .98$). This confirms the robustness of our statistical analyses, showing stable correlations for ρ and τ.

This approach to cognitive plausibility reduces the effectiveness of a similarity or relatedness measure to a single, real number, typically in interval $[-1, 1]$, where 1 means perfect correlation, -1 perfect inverse correlation, and 0 no correlation. The absolute value of such numbers is meaningless, unless it is considered with respect to an upper bound. In similarity research, a typical upper bound is the correlation of the human user closest to the gold standard, likely to be < 1 (e.g. $r = .9$ in [245]). Discussing our Geo Relatedness and Similarity Dataset (GeReSiD), we provide upper bounds both for ρ and τ. As expected, the upper ρ is higher than the upper τ. In this sense, although the *absolute* cognitive plausibility seems different when using different correlation coefficients, the plausibility *relative* to the upper bound is highly consistent.

We think that the cognitive plausibility can be confidently used to assess the effectiveness of semantic similarity and relatedness measures. The next section presents the human-generated dataset we have adopted as a gold standard to run our pilot evaluation.

5.4 MDSM evaluation dataset

This section describes the MDSM evaluation dataset, a set of human rankings of geographic concepts that we utilised as a similarity gold standard to carry out the pilot evaluation of the semantic similarity for OSM. Relevant similarity gold standards were reviewed in Section 2.5.5. This geographic similarity dataset, originally collected by Rodríguez and Egenhofer [250], is suitable to study the cognitive plausibility of co-citation measures. The dataset was utilised to evaluate the Matching-Distance Similarity Measure (MDSM), their semantic similarity measure. They collected similarity judgements for 33 geographic concepts, including large natural entities (e.g. 'mountain' and 'forest'), and man-made features (e.g. 'bridge' and 'house'). Because these concepts were defined in an abstract way through a short lexical definition (without focusing on ontology-specific information), they are suitable to study the cognitive plausibility of our approach.

Judgements were obtained from 72 students through two surveys (A and B), each presenting five questions. Each question consists of a target concept (e.g. 'stadium') and 10 or 11 base concepts to sort according to their similarity to the target. The results indicate the ranking of the concept pairs, from the most to the least similar (e.g. $\langle$athletic field, ball park$\rangle \rightarrow \ldots \rightarrow \langle$athletic field, library$\rangle$). In their evaluation, Rodríguez and Egenhofer [250] focused on the impact of context on similarity judgment. As this aspect of semantic similarity is beyond the scope of this thesis, we excluded from the MDSM evaluation dataset four questions that specify a particular context. Question 5 (target concept: 'lake') is the same in both surveys, therefore it was merged into one question through the mean of the rankings ($QAB5$).

The five questions are non-overlapping, i.e. they strictly contain different pairs. However, an aspect that has to be taken into account in this dataset lies in the high semantic similarity of the target concepts in QA1 and QB1 ('stadium' and 'athletic field'), and in QA4 and QB4 ('path' and 'travelway'). The two sets of questions have been evaluated by different groups of human subjects, and have the purpose of cross-validating the rankings. Hence, a computational measure is expected to perform consistently on both sets of questions. For this reason, all the cognitive plausibilities in the experiments in the remainder of this section are presented separately, as well as aggregated into a meta-analysis to show overall trends.

The 33 concepts of the MDSM dataset were manually mapped onto the corresponding tags in the OSM Semantic Network, based on their textual definitions (see Section 3.3). For example, the concept *tennis court* was matched to `osmwiki:Tag:sport=tennis`. While 29 concepts have a satisfactory equivalent in OSM, four concepts ('terminal,' 'transportation,' 'lagoon,' and 'desert') were discarded because they did not have a precise matching concept in the OSM Semantic Network. The concepts were also manually mapped on WordNet synsets, e.g. 'stadium' corresponds to synset *stadium#n#1*. As a result, we obtained a similarity dataset containing five questions (referred to as $QA1$, $QB1$, $QA4$, $QB4$, $QAB5$), on 29 geographic concepts.

Question	Target concept	Base concepts	Pairs	Tied ranks
$QA1$	stadium	sports arena, ball park, athletic field, tennis court, theater, commons, museum, building, library, house	10	0
$QB1$	athletic field	ball park, sports arena, stadium, tennis court, commons, theater, building, house, museum, library	10	2
$QA4$	travelway	road, highway, path, railway, bridge, subway station, airport, port	8	2
$QB4$	path	road, travelway, bridge, highway, railway, subway station, port, airport	8	0
$QAB5$	lake	forest, city, river, beach, island, wetland, bridge, pond, mountain	9	0

Table 5.1: The MDSM evaluation dataset, based on Rodríguez and Egenhofer [250]. Total pairs: 45

The entire dataset is available online, including the complete manual mapping to OSM and WordNet, and concept definitions.[5] The MDSM evaluation dataset (i.e. the human-generated rankings, not those by the measure MDSM) was used as gold standard to conduct a pilot evaluation for NetLexSiM's network (Section 5.5) and lexical components (Section 5.6).

5.5 NetLexSiM: pilot evaluation of network similarity

In this section we describe an experimental study on the cognitive plausibility of co-citation algorithms, using the MDSM evaluation dataset in the case of the OSM Semantic Network. Co-citation similarity measures constitute the network component of our measure, NetLexSiM. The main goal of this pilot evaluation is to assess the effectiveness of the structural similarity measure outlined in Section 3.6, comparing and contrasting six co-citation measures. In particular, this evaluation has the purpose of validating two hypotheses:

1. Co-citation similarity measures applied to the OSM Semantic Network can reach high cognitive plausibility.
2. The topological structure of the OSM Semantic Network must therefore contain valuable information about geographical concepts.

In the remainder of this section, we outline the experiment setup (Section 5.5.1), and we discuss the experiment results, drawing conclusions on the per-

[5] `http://github.com/ucd-spatial/Datasets`

formance of the six co-citation algorithms (Section 5.5.2). This work has been published in [22].

5.5.1 Experiment setup

To obtain semantic similarity scores for the OSM concepts, we have run several co-citation algorithms on the OSM Semantic Network described in Section 3.6. The underlying intuition is that similar vertices in the graph tend to be linked together, and link the same vertices. P-Rank [318] is a generic algorithm that, with certain combinations of parameters C, K, and λ, is equivalent to Co-citation [271], Coupling [152], Amsler [9], SimRank [136], and rvs-SimRank [318].

Parameter $C \in (0, 1)$ is the decay factor in the propagation of similarity across the graph edges. With high values of C, similarity scores are transferred from a pair of vertices to their neighbours, while low values of C reduce the recursive propagation of similarity. The constant $\lambda \in [0, 1]$ determines the balance between incoming and outgoing links. When λ is equal to 1, the computation of similarity considers incoming links only; on the other hand, when $\lambda = 0$, the computation of similarity is based only on outgoing links. K is the number of iterations through which the similarity scores are computed. When $K = 1$, P-Rank is not iterative. Hence, in order to study the cognitive plausibility of these algorithms, we have computed P-Rank for 550 unique combinations of K, C and λ. The experimental setup is the following (see Table 3.3 for notations):

- λ: 11 discrete equidistant levels $\in [0, 1]$.
- C: 5 discrete equidistant levels $\in [.1, .9]$.
- K: 10 P-Rank iterations.

On top of these 550 combinations, three additional cases were computed: $K = 1$, $\lambda \in \{0, .5, 1\}$ corresponding respectively to the algorithms Co-citation, Coupling, and Amsler. When $K = 1$, the value of C does not influence the results.

Following the approach adopted by Rodríguez and Egenhofer [250], the results were computed on the rankings and not on the similarity scores, i.e. the order of the pairs returned by the system against the order in the MDSM dataset, using Spearman's rank correlation coefficient [276]. Spearman's ρ was computed on each of the five questions, over the 550 combinations. To assess how the algorithms performed overall, a meta-analysis of correlation coefficients had to be carried out for each combination of parameters, across the five questions.

Among the existing meta-analytical methods for correlation coefficients, Field [77] concludes that the "the Hunter-Schmidt method tends to provide the most accurate estimates of the mean population effect size when effect sizes are heterogeneous, which is the most common case in meta-analytic practice" (p.

178). This method was originally developed for Pearson's product moment correlation coefficient [129]. As Altman and Gardner [8] noted, both Pearson's r and Spearman's ρ follow a similar statistical distribution, so that the Hunter-Schmidt method can also be applied in our case.

The aggregated $\bar{\rho}$ is computed through a weighted mean, where the weights are the number of pairs in each question (see Formula 5.1, where q is the number of questions and n_i is the number of pairs in each question). $\bar{\rho}$ expresses the overall correlation between the rankings of P-Rank applied to the OSM Semantic Network, and the MDSM evaluation dataset. To assess the statistical significance of these 550 tests, we have utilised the Hunter-Schmidt method, based on the standard deviation, the standard error, and the Z score [129]. For each set of questions, the standardised Z score is derived from the standard deviation and standard error. The Z score can then be converted to a p-value using the normal distribution table, while the confidence interval can be derived from the standard error [60, p. 523]:

$$\bar{\rho} = \frac{\sum_{i=1}^{q} n_i \rho_i}{\sum_{i=1}^{q} n_i} \qquad SD_{\bar{\rho}} = \sqrt{\frac{\sum_{i=1}^{q} n_i (\rho_i - \bar{\rho})^2}{\sum_{i=1}^{q} n_i}} \qquad (5.1)$$

$$SE_{\bar{\rho}} = \frac{SD_{\bar{\rho}}}{\sqrt{q}} \qquad Z = \frac{\bar{\rho}}{SE_{\bar{\rho}}}$$

For example, given the combination of parameters $\lambda = 1, C = .9, K = 10$, ρ for the five questions are respectively .85, .87, .95, .9, and .7, resulting in $\bar{\rho} = .85$, $SD = .083$, $SE = .037$, $Z = 2326$, therefore $p < .0001$. The 95% confidence interval for $\bar{\rho}$ is $.85 \pm .07$.

5.5.2 Experiment results

The concept rankings for 550 statistically significant cases were generated on the OSM Semantic Network, obtaining correlations with human ranked-pairs of the MDSM dataset. For all of the 550 tie-corrected Spearman correlation tests, we obtained $p < .001$, indicating high statistical significance. Considering only incoming links ($\lambda = 1$), the mean correlation $\bar{\rho}$ is plotted in Figure 5.3. A convergence with $K > 7$ can be observed. The similarity scores fluctuate during the first iterations, and then plateau, remaining stable in the following iterations. As is reasonable to expect, the convergence is more rapid when C is close to 0. Figure 5.4 focuses instead on the parameter λ, showing its impact on the correlation. As λ gets closer to 1, the correlation improves steadily, suggesting that incoming links are more relevant to the computation of similarity than outgoing ones.

The overall impact of the decay factor C is clear, as the best correlations are consistently obtained when $C = .9$, and the worst when $C = .1$ (see Figure 5.3 and 5.4). Given that $C \in (0, 1)$, it is important to look at its impact at the

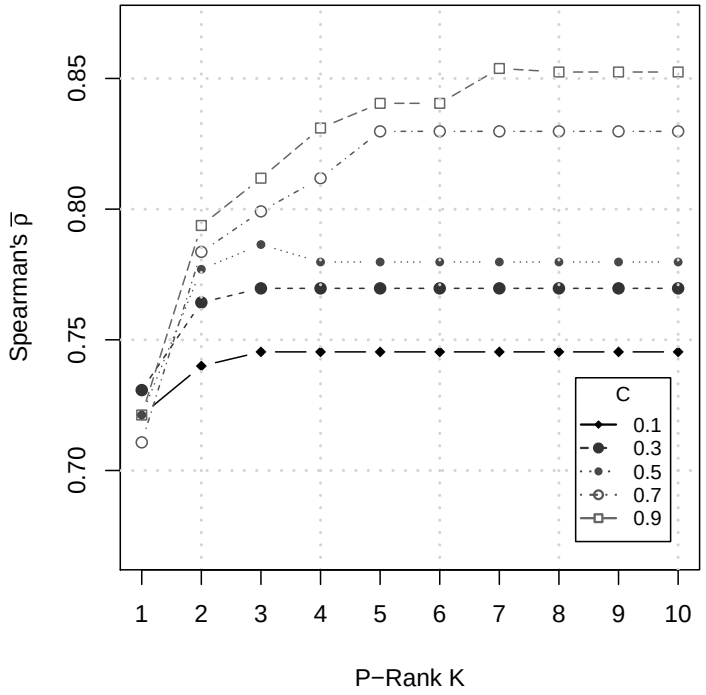

Figure 5.3: Experiment results grouped by C (fixed parameter: $\lambda = 1$); K is the P-Rank iteration, while Spearman's $\bar{\rho}$ is a measure of correlation with human behaviour; $p < .0001$ for all $\bar{\rho}$.

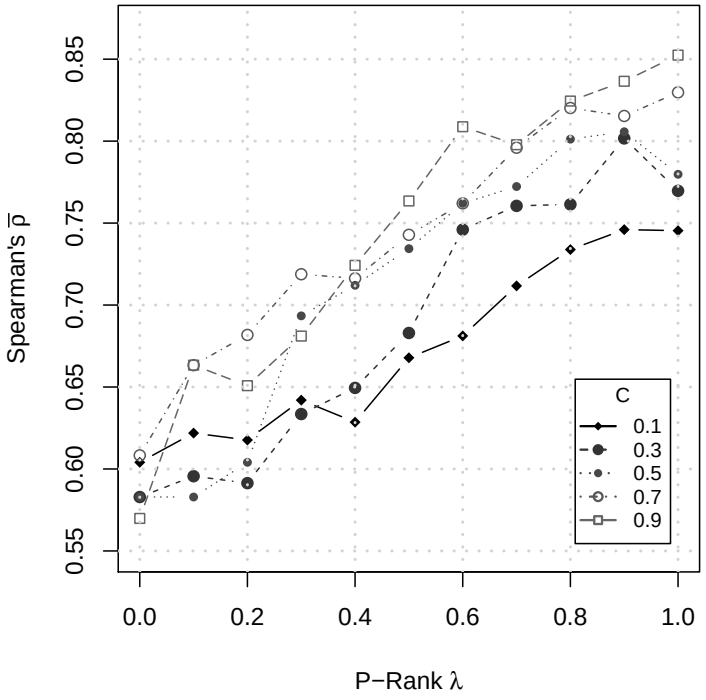

Figure 5.4: Experiment results grouped by C (fixed parameter: $K = 10$); K is the P-Rank iteration, while Spearman's $\bar{\rho}$ is a measure of correlation with human behaviour; $p < .0001$ for all $\bar{\rho}$.

K	λ	C	Algorithm	QA1 ρ	QB1 ρ	QA4 ρ	QB4 ρ	QAB5 ρ	Meta $\bar{\rho}$
			N pairs	10	10	8	8	9	–
1	0	–	Coupling [152]	.57	.71†	.43	.44	.56	.55 $\pm$ .09
1	.5	–	Amsler [9]	.57	.77††	.62	.76†	.64†	.67 $\pm$ .07
1	1	–	Co-citation [271]	.71†††	.77††	.81††	.78†	.55	.72 $\pm$ .08
10	0	.9	rvs-SimRank [318]	.58	.73†	.40	.69†	.42	.57 $\pm$ .12
10	0	.5	–	.58	.73†	.36	.64	.57	.57 $\pm$ .1
10	0	.1	–	.58	.73†	.50	.62	.57	.60 $\pm$ .07
10	.5	.9	P-Rank [318]	.64†	.77†††	.81††	.90††	.73†	.76 $\pm$ .08
10	.5	.5	–	.58	.77†††	.71†	.90††	.73†	.73 $\pm$ .09
10	.5	.1	–	.58	.77†††	.60	.76†	.63	.67 $\pm$ .07
10	1	.9	SimRank [136]	.85††	.87††	.95†††	.90††	.70†	.85 $\pm$ .07*
10	1	.5	–	.85††	.81††	.81†	.86††	.57	.78 $\pm$ .09
10	1	.1	–	.74†	.80††	.81†	.79†	.60	.75 $\pm$ .07
			Question mean	.65	.77	.65	.75	.61	–

Table 5.2: Results of network-based similarity. K, λ and C are the P-Rank parameters. The Spearman rank correlation is the average of the correlations for each of the five questions of the MDSM evaluation dataset; $\bar{\rho}$ is shown with the 95% confidence interval computed with the Hunter-Schmidt method [60]. (*) Best performance. Statistical significance: (†) $p < .05$; (††) $p < .01$; (†††) $p < .001$.

asymptotes. When $C \to 0$, the similarity function $s(a, b)$ tends to $\mathbf{R}_0$. On the other hand, the similarity function does not converge to a finite value when $C \to 1$.

$$\lim_{C \to 0} s(a, b) = \mathbf{R}_0(a, b) \quad \lim_{C \to 1} s(a, b) = \ undetermined \tag{5.2}$$

These properties were confirmed on an additional experiment run with $C \in (.9, 1), K = 100, \lambda = 1$. With $C >= .99$, the similarity scores present strong variations even when $K > 50$, not showing any sign of convergence. According to Jeh and Widom [136], the choice of the optimal value of C depends on the specific domain in which SimRank is being applied. On experimental grounds, we can state that, in the context of the OSM Semantic Network, optimal $C \in [.9, .95]$, which suggests that, to match human judgement, similarity has to flow across the graph edges with a slow decay.

The overall results of the experiment are reported in Table 5.2, which shows the mean Spearman's ρ with 95% confidence intervals, highlighting the overall cognitive plausibility of the algorithms. Among the non-iterative algorithms ($K = 1$), Small's Co-citation performs better than its counterparts. It is possible to notice that, among the iterative algorithms ($K > 1$), SimRank with a low decay ($C = .9$) clearly outperforms the other approaches, reaching a $\bar{\rho} = .85 \pm .07$. Stronger decay factors make the algorithm lose valuable information. The worst results are instead obtained by rvs-SimRank ($\bar{\rho} = .57 \pm .12$), indicating that, in the OSM Semantic Network, outgoing connections between concepts are not strongly correlated to their semantic similarity. This suggests that, when describing concepts in the OSM Wiki, contributors tend to mention

similar concepts together.

On the other hand, citations of the same concept while defining similar classes are statistically less common. For example, considering the links between three OSM tags, *waterway=riverbank* references *waterway=river* and *waterway=stream*, two highly similar concepts. The *waterway=river* tag backlinks *waterway=riverbank*, whilst *waterway=stream* does not. Hence in this case, incoming links from *waterway=riverbank* strengthen similarity between *waterway=river* and *waterway=stream*, while outgoing links do not encode similarity.

As stated in Section 3.3, the OSM Semantic Network contains several types of edges, which are treated equally in this experiment. In order to assess the importance of each edge type, we have run a series of additional experiments including only one type of edge at a time. The co-citation algorithms are not computable when including only sparse edge types such as `osmwiki:key` and `osmwiki:implies`. On the other hand, when including only edges of type `osmwiki:link`, all the algorithms are computable and the corresponding $\bar{\rho}$ are slightly lower than those obtained in the main experiment with all edge types (e.g. $.84 \pm .07$ for SimRank, instead of $.85 \pm .07$). This indicates that the generic hyperlinks `osmwiki:link` convey the bulk of the semantic similarity contained in the network, and the other edges give a minor semantic contribution. Overall, the results show an improvement in the cognitive plausibility as λ moves from 0 to 1, and C from .1 to .9.

The correlation-based approach adopted in this evaluation captures the *overall* cognitive plausibility of a semantic similarity measure. In order to inspect the measures' performance more closely, it is beneficial to discuss failures, i.e. individual concept pairs showing a large discrepancy between the human and the machine-generated rankings. SimRank, the best performing measure, obtains very strong significant correlations with four questions (QA1, QB1, QA4, and QB4), with $\rho > .8$. However, the correlation with question QAB5 is weaker ($\rho = .7$). At close inspection, this lower plausibility is due to the rankings of two concept pairs that show high discrepancy. Out of nine ranked pairs, <*lake, city*> is ranked 6th by human subjects, and 9th by SimRank. While these two concepts are located in different areas of the OSM Semantic Network, and show no structural similarity at all, they might share some similarity in the human subjects' eyes, e.g. many cities are geographically located near lakes.

By contrast, the similarity of <*lake, mountain*> is overestimated by SimRank. This pair is ranked 8th by human subjects, and 4th by SimRank. This discrepancy is caused by the relatively high structural similarity of lake and mountain in the OSM Semantic Network, which are both classified under the key *natural*. Pairs considered more similar by human subjects, e.g. <*lake, city*>, are located under different keys (*natural* and *place*). This mismatch can be mitigated by including the similarity model the semantic similarity of the lexical definitions of terms, as showed in Section 5.6.

Overall, the results outlined in this section show that the SimRank algorithm applied to the OSM Semantic Network closely matches the human judgement in the modified MDSM similarity dataset, reaching the correla-

Question	Target concept	SimRank ρ (dataset: OSN)	MDSM ρ (dataset: WordNet/SDTS)
$QA1$	stadium	.85	.96
$QB1$	athletic field	.87	.92
$QA4$	travelway	.95	.9
$QB4$	path	.9	.88
$QAB5$	lake	.7	.82*
-	-	$\bar{\rho} = .85$	$\bar{\rho} = .89$

Table 5.3: Results for SimRank ($C = .9$, $K = 10$) on the OSM Semantic Network. MDSM results from Rodríguez and Egenhofer [250]. For all correlation tests, $p < .05$. (*) Mean of survey A and B.

tion $\bar{\rho} = .85 \pm .07$ averaged over the five questions. This can be compared with the MDSM evaluation by Rodríguez and Egenhofer [250]. However, the MDSM approach was tested on a geographic ontology derived from the combination of definitions in WordNet and in the Spatial Data Transfer Standard (SDTS), which contain formal knowledge carefully encoded by experts, including parts, functions, and attributes [73, 74]. The network component of NetLexSiM, by contrast, relies exclusively on link patterns in a crowdsourced semantic network of inter-linked geographic concepts, the OSM Semantic Network, ignoring any other meta-information, such as definitions, labels, etc. A comparison between the results of the two approaches is reported in Table 5.3. This confirms our first hypothesis, i.e. co-citation similarity measures applied to the OSM Semantic Network can reach high cognitive plausibility.

These results indicate that, notwithstanding the lack of rich formal semantics in OSM, it is possible to extract a plausible semantic similarity measure from its crowdsourced semantic network, matching closely the performance obtained on a knowledge-rich formal structure such as WordNet and Spatial Data Transfer Standard (SDTS). In this case, a crowdsourced resource, such as the OSM Semantic Network, closely matches expert-authored knowledge bases. Moreover, this confirms our second hypothesis, stating that the topological structure of the OSM Semantic Network contains valuable information about geographical concepts. Based on the collected evidence, we draw the following conclusions:

- SimRank [136] applied to the OSM Semantic Network offers a viable semantic similarity measure for the network component of NetLexSiM, obtaining a correlation $\bar{\rho} = .85$.
- Among the non-iterative algorithms, classic Co-citation [271] obtains relatively high cognitive plausibility ($\bar{\rho} = .72$).

This evaluation will be extended and conducted on our dataset GeReSiD in Section 6.3, obtaining highly consistent results. The next section reports the pilot evaluation of the second component of NetLexSiM, the lexical semantic similarity of concept definitions, expressed in natural language on the OSM Wiki website.

5.6 NetLexSiM: pilot evaluation of lexical similarity

This section describes a pilot evaluation for the second component of NetLexSiM, the semantic similarity technique outlined in Section 3.7. The purpose of this evaluation is to assess the cognitive plausibility of our approach, using the MDSM evaluation dataset as a gold standard. Given two definitions of geographic concepts a and b, we extract semantic terms. These terms are combined in semantic vectors representing the concepts. Finally, the similarity is quantified by comparing the semantic vectors. The recursive intuition in which this approach is grounded is that similar terms tend to be defined using similar terms.

To explore the performance of our approach, we computed the semantic similarity of a set of pairs for a number of parameter combinations, and we compared the resulting rankings with a human-generated dataset, the MDSM evaluation dataset (see Section 5.4). This evaluation permits the detailed observation of the impact of parameters, part-of-speech (POS) tags, and term-to-term similarity measures on the final concept-to-concept scores, providing guidelines on what techniques offer the most cognitively plausible measure. The hypotheses that this pilot evaluation wants to validate are the following:

1. The lexical definitions of OSM concepts allow the computation of a more plausible semantic similarity measure than the simple tags.
2. Our bag-of-words (BOW), WordNet-based approach can reach a high cognitive plausibility.

To validate the first hypothesis, i.e. the OSM simple tags are insufficient to compute plausible semantic similarity, we ran a preliminary experiment. Using the manual mapping between the 29 concepts in the MDSM evaluation dataset and the corresponding WordNet synsets, similarity scores were computed for the ten WordNet-based similarity measures outlined in Section 2.5.4. The tie-corrected Spearman's ρ was then obtained for each measure. The cognitive plausibility of this approach turned out to be poor, with ρ falling in the interval $[.24, .5]$. Moreover, for several measures, given the small size of each of the five questions in the gold standard, the statistical significance of the correlation test was largely insufficient ($p > .1$). The inadequacy of this approach validates the hypothesis, supporting the necessity of including the lexical definitions in the similarity computation.

Another popular family of approaches to similarity that does not obtain satisfactory results in our context, is that based on set-theoretical term overlap. A term is similar to another term to the degree to which they share the same terms in their lexical definition, seen as a bag-of-words (BOW), using Tversky or related set-theoretical approaches, such as the Dice coefficient [295, 44]. To evaluate these approaches, another preliminary experiment was set up. The lexical similarity of the five questions of the MDSM evaluation dataset was computed using the Text::similarity tool by Ted Pedersen on the

OSM definitions.[6] As the lexical definitions share a very limited number of terms, mostly generic terms such as 'refer' or 'tag', these techniques incur a limited-information problem, resulting in a set of null similarities (43 out of the 47 scores were equal to 0). These similarity scores showed no correlation with the human scores ($\rho \in [-.1, .1]$ for all the set-theoretical measures), confirming the need for a more sophisticated approach, going beyond the simple boolean, set-theoretical comparison between terms.

The setup of the pilot experiment for our WordNet-based lexical similarity measure is outlined in Section 5.6.1, while the experiment execution is described in Section 5.6.2. Section 5.6.3 discusses in detail the experiment results.

5.6.1 Experiment setup

Our approach to concept-to-concept lexical similarity consists of comparing concept lexical definitions extracted from the OSM Semantic Network. The four parameters that have an impact on the algorithms are the following:

1. **POS filter** (POS). When building the vectors, only terms t having certain part-of-speech tags (POS) are included. Two POS tags are considered in this experiment: nouns (*NN*), and verbs (*VB*). Adjectives (*JJ*) were initially included, but most measures sim_t are designed to handle only nouns and verbs, making a direct comparison difficult. For this reason we excluded adjectives from the experiment, leaving them to future work.

2. **Corpus** (C). The semantic weights in the vectors are determined with a statistical weighting mechanism, i.e. Term Frequency-Inverse Document Frequency (TF-IDF), based on a text corpus C. Different corpora determine different vectors. We collected two corpora for this experiment: the OSM Wiki website,[7] and a set of random news stories from the newspaper *Irish Independent*.[8] The OSM Wiki website corpus is strongly skewed towards geographic and OSM-related terms. On the other hand, the Irish Independent has the purpose of representing an example of a non-geographic corpus. In order to keep the two corpora comparable, the size of the Irish Independent corpus was limited to about 5,000 documents. Table 5.5 summarises the text corpora.

3. **Term-to-term Similarity** (sim_t). The term-to-term similarity function sim_t is utilised by the vector-to-vector similarity to compare two semantic vectors. Given the vast literature covering WordNet as a tool to compute semantic similarity, we exploit WordNet-based semantic similarity measures, in particular the work of Pedersen et al. [234]. Table

[6] `http://text-similarity.sourceforge.net`
[7] `http://wiki.openstreetmap.org`
[8] `http://www.independent.ie`

Symbol	#	Description
POS	3	POS filters combinations (*NN*, *VB*, *NN VB*)
C	3	Text corpora: Null, OSM Wiki, Irish Independent (see Table 5.5)
sim_t	10	Term-to-term similarity measures: *hso*, *jcn*, *lch*, *lesk*, *lin*, *path*, *res*, *vector*, *vectorp*, *wup* (see Table 5.6)
sim_v	2	Vector-to-vector similarity measure: *com* (Corley and Mihalcea [54]), and *fes* (Fernando and Stevenson [75])
Total	180	$\|POS\| \cdot \|C\| \cdot \|sim_t\| \cdot \|sim_v\|$

Table 5.4: Lexical similarity evaluation: resources included in the experiment

Corpus Name	Extracted on	Doc #	Term #	Description
OSM Wiki website	Oct 13, 2011	5,407	2.49M	Official wiki website of the OSM community.
Irish Independent	May 28, 2011	4,868	2.18M	Random news stories from a daily newspaper.
Null	–	–	–	Constant weight assigned to all terms in a vector.

Table 5.5: Text corpora C, used to weight terms in semantic vectors.

5.6 shows the ten semantic similarity measures included in the experiment. The information content information used in the experiment was extracted from the English SemCor corpus, widely used in combination with WordNet to compute semantic similarity [207].

4. **Vector-to-vector Similarity** (sim_v). The similarity function sim_v compares two semantic vectors exploiting the term-to-term function sim_t. We include the measures *com* (Corley and Mihalcea [54]), and *fes* (Fernando and Stevenson [75]), originally developed to detect paraphrases.

In order to explore in detail the performance of our approach to lexical similarity of geographic concepts, we have included several options for the four parameters $\{POS, C, sim_t, sim_v\}$: three POS combinations, ten WordNet-based term-to-term similarity measures (see Section 2.5.4), three text corpora, and two vector-to-vector measures. These options are summarised in Table 5.4.

The pilot experiment consisted of the computation of the scores of the 47 concept pairs contained in the MDSM evaluation dataset, for the 180 combinations of parameters. A combination of parameters $\{POS, C, sim_t, sim_v\}$ returns a unique set of similarity scores for the concept pairs. The cognitive plausibility of these scores was then compared with the MDSM evaluation dataset through tie-corrected Spearman ρ, and a meta-analysis to combine the five questions into a single value (see Section 5.5.1).

5.6.2 Experiment pre-processing

The lexical definitions of the 29 concepts were extracted from the OSM Semantic Network. Subsequently, the definitions were lemmatised with Stan-

Symbol	Type	Reference	Approach
com	sim_v	Corley and Mihalcea [54]	Sum of maximum similarity scores between terms
fes	sim_v	Fernando and Stevenson [75]	Sum of all similarity scores between terms
hso	sim_t	Hirst and St-Onge [124]	Paths in lexical chains
jcn	sim_t	Jiang and Conrath [137]	Information content of lcs and terms
lch	sim_t	Leacock and Chodorow [170]	Edge count scaled by depth
lesk	sim_t	Banerjee and Pedersen [28]	Extended gloss overlap
lin	sim_t	Lin [180]	Ratio of information content of lcs and terms
path	sim_t	Rada et al. [239]	Edge count in shortest path between terms
res	sim_t	Resnik [245]	Information content of lcs of the terms
vector	sim_t	Patwardhan and Pedersen [232]	Second order co-occurrence vectors
vectorp	sim_t	Patwardhan and Pedersen [232]	Pairwise second order co-occurrence vectors
wup	sim_t	Wu and Palmer [313]	Edge count between lcs and terms

Table 5.6: Summary of the relatedness/similarity measures included in the experiment. lcs stands for lowest common subsumer. See Section 2.5.4 for a detailed description of these measures.

ford Core NLP,[9] and POS-tagged with the Stanford Log-linear Part-Of-Speech Tagger [290]. Of all the tagged terms, only nouns, verbs, and adjectives were selected (respectively *NN*, *VB*, and *JJ*). As a result, 1,406 terms were selected, of which 789 were nouns, 236 adjectives, and 381 verbs. Because of the difficulties of computing their similarity in WordNet, adjectives were finally discarded.

Each of the 29 concepts is therefore described by a set of terms. For example, the concept *athletic field* (OSM tag *sport=athletics*) is defined as: "Track and field athletics. A collection of sports events that involve running, throwing and jumping."[10] From this definition the following set of terms is extracted:

track/NN field/NN athletics/NN collection/NN sport/NN event/NN involve/VB run/VB throw/VB jump/VB

The size of these 29 sets varies greatly, depending on the length of the definitions on the OSM Wiki website. Figure 5.5 shows the distribution of size over the 29 sets. It is possible to notice that most definitions have between 20 and 50 terms, with a tail of large cases.

Once the sets are ready, the semantic vectors are constructed by computing semantic weights for each term, from a corpus C. In this experiment we chose classic TF-IDF as the term weighting scheme [116]. In the case of the *Null* corpus, all the terms have the same semantic weight. To illustrate this point,

[9] `http://nlp.stanford.edu/software/corenlp.shtml`
[10] `http://wiki.openstreetmap.org/wiki/Tag:sport%3Dathletics`

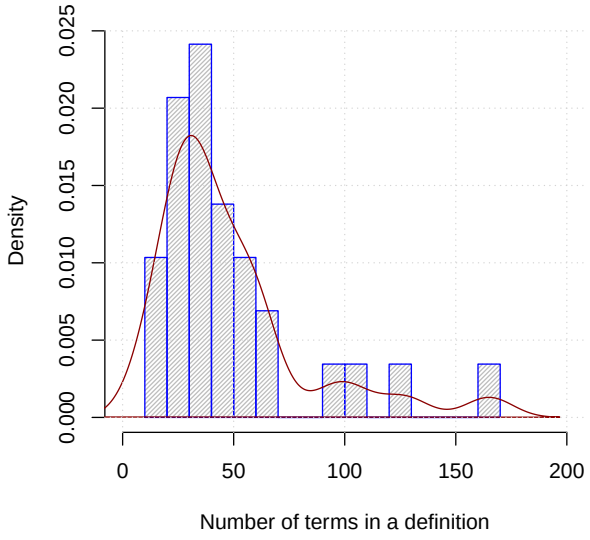

Figure 5.5: Distribution of the size of concept definitions. $Mean = 48.48$, $SD = 34.87$

Semantic vector term	OSM Wiki Website	Irish Independent	Null Corpus
track/NN	.0654	.1022	.1
field/NN	.0870	.1069	.1
athletics/NN	.1714	.2042	.1
collection/NN	.1084	.1147	.1
sport/NN	.0753	.0954	.1
event/NN	.1241	.0669	.1
involve/VB	.1168	.0593	.1
run/VB	.0788	.0496	.1
throw/VB	.1530	.0930	.1
jump/VB	.0193	.1073	.1

Table 5.7: Semantic vectors for concept *athletic field*, with the semantic weights w for the three corpora included in the experiment. $\sum w = 1$

Table 5.7 shows the semantic vectors constructed for *sport=athletics*, for each of the three corpora included in the study. For example, the noun 'event' is a lot less frequent in the OSM Wiki website than in the Irish Independent corpus, resulting in a higher weight.

In order to compare the vectors through the *com* and *fes* and vector-to-vector measures, term-to-term scores have to be computed. The open source project WordNet::Similarity,[11] managed by Ted Pedersen, implements all of these measures of semantic similarity on WordNet, and was used in this phase of the evaluation [234]. Using WordNet::Similarity, we pre-computed the complete similarity matrices for all the term-term measures. In theory, for the 1,406 terms included in this study, 988,418 term pairs needed to be pre-computed ($\frac{n^2}{2}$). To compute the similarity of a specific term pair, more computations are

[11] `http://wn-similarity.sourceforge.net`

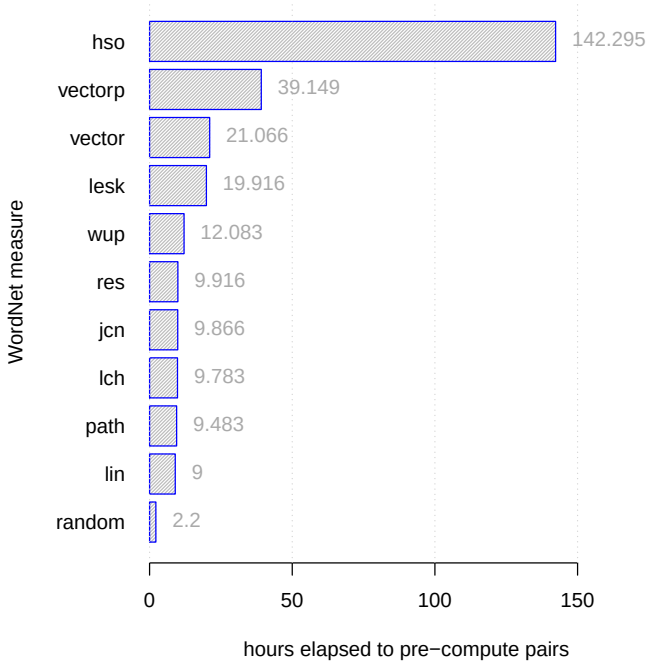

Figure 5.6: Computing time necessary to pre-compute 552,000 term pairs through WordNet::similarity (total: 73.1M pairs). The computation was carried out on a Dell XPS 8300 machine (3.0GHz, Linux Ubuntu 10.10). The random measure is displayed as a base line.

needed. As word senses are not disambiguated, we adopt as a simple heuristics the selection of the maximum similarity score among all the combinations of word senses (see Section 2.5.4). Word sense disambiguation is a complex research problem, and is outside the scope of this study [220].

When comparing nouns 'field' (17 senses) and 'area' (6 senses), 102 pairs have to be computed, and the maximum score obtained is selected. For this reason, for each term-to-term measure sim_t, 552,000 pairs had to be pre-computed. This pre-computation offers the possibility of empirically observing the temporal complexity of each measure. Figure 5.6 shows the computing time necessary to obtain similarity scores for the all the pairs. Bearing in mind the variability induced by the implementation of the measures, it is possible to notice that shortest path-based measures (*path*, *res*, *jcn*, *lin*, and *wup*) are remarkably faster to compute than gloss-based ones (*lesk*, *vector*, and *vectorp*). The *hso* measure is substantially more complex than the others, as the construction of the lexical chains described by Hirst and St-Onge [124] requires visiting several paths between two terms. However, considering that these measures can only compute the similarity of terms belonging to the same POS (noun, verb or adjective), several pairs were discarded, obtaining a total of approximately 552,000 pairs.

Having computed this set of scores, it is possible to analyse it as a sample of map-related concept pairs from WordNet. This sample of similarity scores can highlight the mutual correlation of the ten measures, highlighting their

	hso	jcn	lch	lesk	lin	path	res	vectorp	vector	wup
hso	1.00	–	–	–	–	–	–	–	–	–
jcn	.12	1.00	–	–	–	–	–	–	–	–
lch	.25	.32	1.00	–	–	–	–	–	–	–
lesk	.1	.29	-.08	1.00	–	–	–	–	–	–
lin	.25	.43	.12	.29	1.00	–	–	–	–	–
path	.21	.29	.92	-.26	-.04	1.00	–	–	–	–
res	.31	.04	.1	.28	.68	-.12	1.00	–	–	–
vectorp	.05	.16	.1	.15	.05	.12	.0	1.00	–	–
vector	.09	.43	.19	.53	.17	.15	.03	.43	1.00	–
wup	.27	.24	.7	.21	.49	.44	.7	.04	.16	1.00

Table 5.8: Correlation matrix of the ten WordNet-based measures, on a sample of 552,000 similarity scores.

properties. To investigate how different the measures are on an empirical basis, we compute the Spearman's correlation for the $\frac{n^2-n}{2}$ (45 with $n = 10$) pairs of measures, comparing directly the similarity scores in the sample. The resulting 10×10 symmetric correlation matrix is depicted in Table 5.8. All the correlation tests obtain high statistical significance ($p < .0001$). The correlation coefficients ρ of the WordNet measures fall in the interval $[-.26, .92]$, indicating a high variability. A few pairs show strong correlation: *(path,lch) (wup,lch) (wup,res)*, $\rho \in [.7, .92]$. A wider group displays a medium correlation: *(lin,jcn) (res,lin) (vector,jcn) (vector,lesk) (vector,vectorp) (wup,lin) (wup,path)*, with $\rho \in [.4, .7)$. All the other pairs in the matrix show weak or no correlation $\rho \in (.4, -.26]$.

These results can be partly accounted for by observing the specific approaches to similarity, summarised in Section 2.5.4. A family of measures relies exclusively on the shortest path in the taxonomy (*path*, *lch*, *wup*), whilst other measures also include information content (*res*, *lin*, *jcn*). Analogously, gloss-based measures can be clustered together (*lesk*, *vector*, *vectorp*). By contrast, the measure *hso*, based on a different approach (i.e. lexical chains between terms), shows little correlation with the other measures. However, even in the same family of measures, some measures show divergent behaviour. For example, *res* and *jcn* are conceptually similar, i.e. both rely on shortest path and information content, but their mutual ρ is only $.04$. This variability is reflected in the final results, in which some measures obtain high cognitive plausibility, and others low, without any family of measures showing a clear dominance over the others.

5.6.3 Experiment results

This section shows the results obtained from the pilot evaluation on the cognitive plausibility of our approach to lexical similarity.

For each of the 180 combinations $\{POS, C, sim_t, sim_v\}$, we computed the tie-corrected Spearman's correlation coefficient ρ. The correlation ρ is computed for each of the five questions in the MDSM evaluation dataset. To assess how the similarity measures performed overall, a meta-analysis of correlation coefficients had to be carried out for each combination of parameters, across

Param Name	Param Value	median $\tilde{\bar{\rho}}$	25%-75% quartiles		max $\bar{\rho}$
sim_v	com	.7*	.64	.76	.81
	fes	.66	.56	.74	.79
POS	NN VB	.7*	.61	.75	.81
	NN	.68	.57	.74	.81
	VB	–	–	–	–
C	Null	.7*	.62	.76	.81
	Irish Indep	.7	.6	.74	.79
	OSM Wiki	.67	.59	.74	.81
sim_t	jcn	.75*	.73	.75	.77
	lch	.74	.69	.76	.81
	wup	.74	.66	.78	.8
	res	.72	.69	.77	.81
	hso	.71	.7	.73	.76
	path	.7	.66	.76	.77
	lin	.67	.58	.75	.81
	vector	.62	.58	.66	.68
	vectorp	.57	.55	.64	.66
	lesk	.48	.44	.55	.56
Question	A1	.58	.48	.65	.92
	B1	.76	.68	.88	1
	A4	.71	.52	.82	.95
	B4	.52	.24	.73	.95
	AB5	.72	.59	.82	.95
All	–	.69	.6	.64	.81

Table 5.9: Summary of results of the lexical similarity experiment. The central tendency of each parameter is summarised by the median $\tilde{\bar{\rho}}$, its lower and upper quartiles, and its maximum value. The distribution of ρ is also showed for each question of the MDSM evaluation dataset. (*) Best performance.

the five questions, using the Hunter-Schmidt meta-analysis method [77].

The method applied in this experiment is analogous to that used in the evaluation of the network-based similarity in Section 5.5. The resulting average $\bar{\rho}$ expresses the overall correlation between the rankings computed by our similarity approach applied on the OSM Semantic Network, and the MDSM evaluation dataset. For the cases including only verbs (POS=VB), the correlations with the human dataset were very weak ($\bar{\rho} \in [-.1, .4]$), and obtained insufficient statistical significance ($p > .1$). For this reason, we excluded these cases from the analysis, and the resulting correlations are not reported. By contrast, all the other parameter combinations obtained high statistical significance ($p < .001$).

To measure the performance of a particular parameter (e.g. $C = Null$ or $sim_t = lin$), all the cases where that parameter is used are treated as a set of $\bar{\rho}$. By looking at the distribution of $\bar{\rho}$ within these sets, it is possible to notice that the distributions are often skewed to the left, i.e. towards lower values. For this reason, we consider the median to be a more robust descriptor of central tendency than the mean.

Table 5.9 reports the detailed results of the cognitive plausibility of our

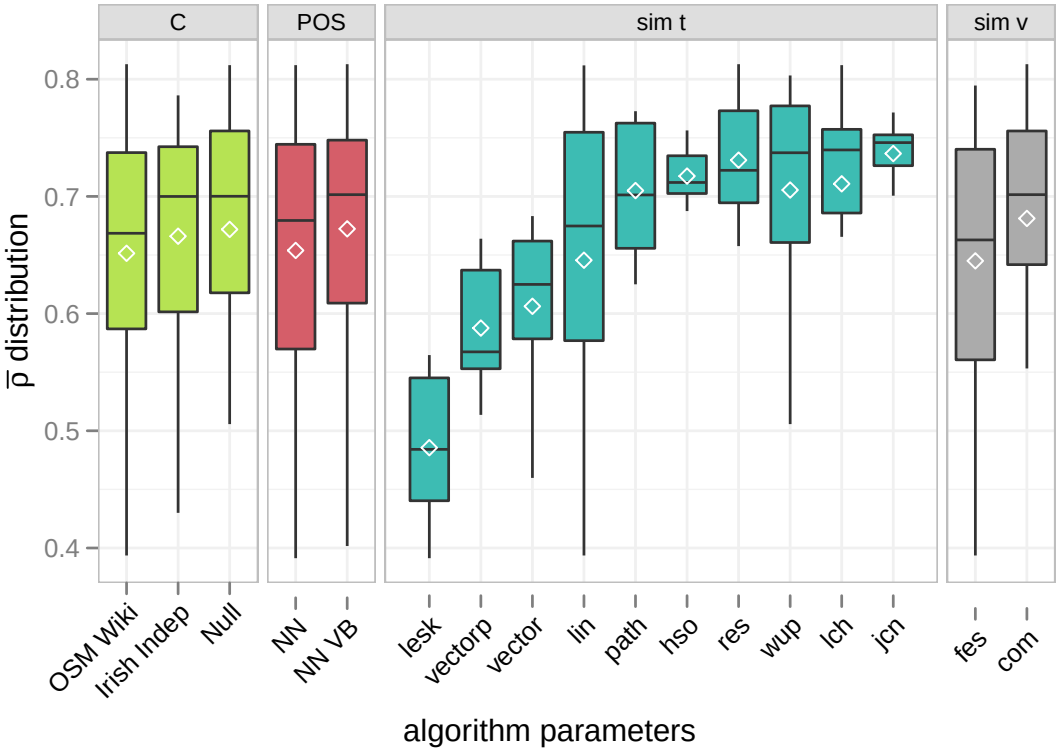

Figure 5.7: Lexical NetLexSiM pilot experiment: cognitive plausibility against the MDSM evaluation dataset. The boxplot shows the distribution of $\bar{\rho}$, sorted by the median in ascending order. For each value for the parameter, the box plot shows the smallest $\bar{\rho}$, the 25% quartile, the median, the 75% quartile, and largest $\bar{\rho}$. The white diamond represents the mean.

approach to lexical similarity. For each parameter, the table reports the median, the 25% and 75% quartiles, and the maximum value, which give a clear picture of the experiment outcome for a specific parameter option. For example, when the POS filter is set to *NN*, the distribution of $\bar{\rho}$ results in a $\tilde{\bar{\rho}} = .68$, with a maximum value of .81. The distributions of $\bar{\rho}$ are displayed in Figure 5.7. Each boxplot indicates six points of the distribution: minimum value, 25% quartile, median, 75% quartiles, maximum, and mean (rendered as a white diamond on the boxplots). This allows an intuitive understanding of the impact of each parameter option on the $\bar{\rho}$ distribution.

The inclusion of certain part-of-speech tags (POS) have a noticeable impact on the results. As is reasonable to expect given the descriptive nature of the OSM Semantic Network definitions, nouns (*NN*) carry most of the concept semantics, with $\tilde{\bar{\rho}} = .68$. Verbs (*VB*) showed a very weak correlation with the human dataset, and because of the low statistical significance of their correlation tests were not included in the analysis. The combination of nouns and verbs (*NN VB*) performs better than nouns by themselves ($\tilde{\bar{\rho}} = .7$). The inclusion of verbs also slightly reduces the dispersion, and the best results are obtained when nouns and verbs are both included.

The corpus C determines the semantic weights in the vectors. Figure 5.7

shows $\bar{\rho}$ grouped by the three corpora included in this study. The Null and the Irish Independent corpora have very close medians (.7), with the quartiles of the distribution of Null being higher. The OSM Wiki website shows slightly lower cognitive plausibility ($\tilde{\bar{\rho}} = .67$). Therefore, it is clear that the current weighting scheme (TF-IDF) does not improve the cognitive plausibility of the approach. This is consistent with the findings of Fernando and Stevenson [75], who showed that in the area of paraphrase detection such weighting schemes slightly *worsen* the performance.

The reasons for this counter-intuitive result can be various. The selected corpora, OSM Wiki website and the Irish Independent corpus, might be too small or biased. Similarly, the weighting scheme (classic TF-IDF [140]) could be sub-optimal compared to other schemes, such as BM25 [141]. However, in the evaluation on our gold standard GeReSiD, the impact of C obtains different results, suggesting that the Irish Independent corpus considerably improves the results over the other corpora (see Section 6.4).

The vector similarity measure sim_v has a strong impact on the approach performance. Corley and Mihalcea's *com* approach shows high cognitive plausibility ($\tilde{\bar{\rho}}= .7$), while Fernando and Stevenson's *fes* tends to obtain lower results ($\tilde{\bar{\rho}}= .66$). While in paraphrase detection *fes* performs slightly better than *com*, in context of OSM semantic similarity the opposite is true [75]. In general, the measure *com*, defined by Corley and Mihalcea [54], should be preferred in the computation of the semantic similarity of geographic concepts.

Ultimately, the purpose of this study is to provide guidelines on what technique offers the most cognitively plausible measure of semantic similarity for OSM geographic concepts. To reach this goal, it is useful to observe in detail the performance of our approach when selecting top performing parameters. While three parameters show clear trends and have optimal values (POS=*NN VB*, C=*Null*, sim_v=*com*), the term-to-term measure sim_t displays higher variability, and its discussion requires more caution.

The measure *jcn* has the highest mean (.75), and reaches .77 as max $\bar{\rho}$. Measures *lch* and *wup* have a comparable median (.74), and obtain a higher max value ($\approx .8$). However, as it is possible to notice in Figure 5.7, *wup* has a wider distribution, with a minimum substantially lower than *jcn* and *lch* (.51). Another measure that falls in the top performing group is *res*, with a lower median (.72), but relatively high quartiles and maximum (respectively .69, .77, and .81). Thus, the three measures that provide the most promising results are *jcn* (Jiang and Conrath [137]), *lch* (Leacock and Chodorow [170]), and *res* (Resnik [245]). All the other measures have either considerably lower medians, maximum values, or have very wide distributions that show low stability in the results.

Fixing the other parameters to optimal values, the cognitive plausibility of the three top performing sim_t is $\bar{\rho} = .77 \pm .23$ for *jcn*, $.76 \pm .17$ for *lch*, and $.77 \pm .1$ for *res*. These three measures maintain a high performance even with suboptimal POS (*NN*), respectively $.76 \pm .18$ (*jcn*), $.81 \pm .09$ (*lch*), and $.7 \pm .08$ (*res*). The stability of these results is further confirmed when selecting a suboptimal corpus C, such as the OSM Wiki website. Even in this case, the correlation

Parameters				Question ρ					Meta
sim_v	POS	C	sim_t	A1	B1	A4	B4	AB5	$\bar{\rho}$
com	NN	null	lch	$.75^{\dagger}$	$.92^{\dagger\dagger\dagger}$	$.79^{\dagger}$	$.86^{\dagger}$	$.76^{\dagger}$	$.81 \pm .09^{\dagger\dagger\dagger}$
com	NN VB	null	lch	.59	$.95^{\dagger\dagger\dagger}$	.67	$.88^{\dagger\dagger}$	$.73^{\dagger}$	$.76 \pm .17^{\dagger\dagger\dagger}$
com	NN	null	jcn	.55	$.91^{\dagger\dagger\dagger}$	$.81^{\dagger}$	.52	$.84^{\dagger\dagger}$	$.76 \pm .18^{\dagger\dagger\dagger}$
com	NN VB	null	jcn	.49	$.95^{\dagger\dagger\dagger}$	$.93^{\dagger\dagger}$	.55	$.89^{\dagger\dagger}$	$.77 \pm .23^{\dagger\dagger\dagger}$
com	NN	null	wup	$.66^{\dagger}$	$.92^{\dagger\dagger\dagger}$	.69	.67	$.74^{\dagger}$	$.78 \pm .13^{\dagger\dagger\dagger}$
com	NN VB	null	wup	.54	$.95^{\dagger\dagger\dagger}$	.57	$.86^{\dagger}$	$.70^{\dagger}$	$.73 \pm .19^{\dagger\dagger\dagger}$
com	NN	null	res	$.77^{\dagger}$	$.74^{\dagger}$	.57	.67	.59	$.70 \pm .08^{\dagger\dagger\dagger}$
com	NN VB	null	res	$.67^{\dagger}$	$.89^{\dagger\dagger\dagger}$	.57	.64	$.74^{\dagger}$	$.77 \pm .10^{\dagger\dagger\dagger}$
		Question	mean	.64	.89	.72	.69	.76	–

Table 5.10: Details of best cases. Statistical significance: (†) $p < .05$; (††) $p < .01$; (†††) $p < .001$.

with the human dataset falls in the interval $[.7, .81]$. Details of the best cases are reported in Table 5.10, including the correlations for the separate questions.

These results validate our approach to lexical similarity, informed by WordNet-based, BOW techniques. Our two hypotheses are confirmed: the lexical definitions of OSM concepts allow the computation of a more plausible semantic similarity measure than the simple tags, and our technique, based on WordNet and paraphrase detection, can reach a high cognitive plausibility. While the OSM tags do not suffice to compute semantic similarity, the lexical definitions contained in the OSM Semantic Network provide a useful semantic resource. From a pragmatic viewpoint, in order to obtain the highest cognitive plausibility with the lexical component of NetLexSiM, the following policies should be adopted:

- **POS filter**: include nouns; verbs slightly improve the results.
- **Corpus** C: apply uniform weights; TF-IDF weighting scheme does not improve the results.
- **Vector-to-vector similarity measure** sim_v: adopt the paraphrase detection technique devised by Corley and Mihalcea [54].
- **Term-to-term WordNet-based similarity measure** sim_t: adopt a technique either by Jiang and Conrath [137], Leacock and Chodorow [170], or Resnik [245].

Using these optimal parameters, the cognitive plausibility indicator $\bar{\rho}$ falls in the interval $[.7, .81]$, showing a strong correlation with the human-generated MDSM evaluation dataset. Even without a rich formalisation such as description logic (DL), it is possible to obtain a highly plausible measure of semantic similarity of volunteered geographic concepts, purely relying on lexical definitions.

Unlike SimRank, whose cognitive plausibility is discussed in Section 5.5, the plausibility of the lexical similarity measures varies considerably across the five questions of the MDSM evaluation dataset. On average, while questions QA4, QB1, and QAB5 tend to obtain strong correlations ($\rho > .7$), QA1 and QB4

show lower plausibility ($\rho < .7$). In particular, the target concept in question QB4 (*path*) has a long and complex lexical definition, which the algorithm fails to handle. Many of the 67 nouns contained in this lexical definition are related to the definition of path, and discuss specific issues relevant for OSM. Nouns that are used to differentiate a path from similar concepts, such as *motor*, *vehicle*, *navigation*, *restriction*, and *access*, decrease the representativeness of the resulting semantic vector.

Analogous issues occur in QA1, whose target concept is *stadium*, while the target concepts of the other three questions lead to better semantic vectors (respectively *athletic field*, *travelway*, and *lake*). A possible solution to this issue might consist of increasing the weight of the semantic terms at the beginning of lexical definitions. In the aforementioned cases, the noise tends to be located at the end of the lexical definition, where minute details are discussed, while the first terms in the definition seem to be more semantically important. As shown in Chapter 6, these limitations are overcome by integrating the lexical similarity with the structural similarity of the geographic concepts.

The results described in this section ($\bar{\rho} \in [.7, .81]$) are lower than those presented by Rodríguez and Egenhofer [250] ($\bar{\rho} = .89$). However, it is important to point out that the two evaluations were conducted on completely different datasets, adopting unrelated approaches to semantic similarity. Rodríguez and Egenhofer [250] utilised their similarity measure, the Matching-Distance Similarity Measure (MDSM), on a geographic ontology based on the combination of WordNet and SDTS, a knowledge-rich topological geospatial standard. By contrast, we applied our lexical similarity measure only on the crowdsourced natural-language lexical definitions from the OSM Wiki website, which are inherently noisy, inconsistent, of variable quality, clarity and length. Taking this crucial difference into account, we consider our results to be promising.

In this section we have reported a pilot evaluation for the lexical component of NetLexSiM. This experiment explored several measures of semantic similarity, including ten existing WordNet-based measures in the process. A similarity measure is never simply right or wrong, but rather offers an approximation of human judgement, reaching a degree of cognitive plausibility in a given context. As discussed earlier, some term-to-term measures sim_t obtain comparable results, and yet show visible differences in their behaviour. In this sense, a similarity measure can be seen as a human expert giving a judgement on a complex problem, and disagreeing with her peers. When critical decisions have to be taken, a typical approach to reach a reliable conclusion, consists of consulting a number experts. The next section expands this intuition, introducing the concept of the 'similarity jury.'

5.7 The similarity jury

This section presents the initial exploration of the combination of multiple term-to-term similarity measures, conducted as part of the pilot evaluation of

the lexical component of NetLexSiM, our similarity measure for OSM concepts (see Section 5.6). After defining the 'similarity jury,' we describe an experiment involving eight similarity measures, whose cognitive plausibility is computed using the MDSM evaluation dataset. Subsequently we discuss the results of this experiment, and we outline pros and cons of the jury with respect to individual measures. This work was published in [24].

Our investigation of the computation of lexical similarity of OSM concepts includes a number of term-to-term similarity measures sim_t (see Section 5.6). Given the wide variety of existing lexical similarity techniques sim_t, choosing a suitable measure in a new context is a significant issue. Approaches to lexical similarity are not intrinsically right or wrong, but obtain a certain degree of cognitive plausibility within a given context. When a specific gold standard is available, one reasonable policy is to select the measure showing the highest correlation with it. However, such gold standards are difficult to construct and validate, and the choice of the most appropriate measure remains highly problematic in many contexts. A semantic similarity measure can be seen as a human domain expert summoned to rank pairs of concepts, according to her subjective set of beliefs, perceptions, hypotheses, and epistemic biases.

In this study, we identify an analogy between computable semantic similarity measures and the combination of expert judgements, a problem relevant to several areas. Indeed, expert disagreement is not an exceptional state of affairs, but rather the norm in human activities characterised by uncertainty, complexity, and trade-offs between multiple criteria [213]. As Mumpower and Stewart [216] put it, the "character and fallibilities of the human judgement process itself lead to persistent disagreements even among competent, honest, and disinterested experts" (p. 191).

Because of the high uncertainty in complex systems, experts often disagree on risk assessment, infrastructure management, and policy analysis [213, 53]. Mathematical and behavioural models have been devised to elicit judgements from experts for risk analysis, suggesting that simple mathematical methods perform quite well [48]. From a psychological perspective, in cases of high uncertainty and risk (e.g. choosing medical treatments, long term investments, etc), decision makers consult multiple experts, and try to obtain a representative average of divergent expert judgements [40].

To the best of our knowledge, there are no studies that address the possibility of combining lexical similarity measures in the context of geographic concepts. In the remainder of this section, we formalise the idea of the jury (Section 5.7.1), we describe an experiment for its evaluation (Section 5.7.2), and we discuss the results, drawing conclusions about the similarity jury (Section 5.7.3).

5.7.1 The jury

A computable measure of semantic similarity can be seen as a domain expert who ranks pairs of concepts. When a gold standard is available, it is possible

to measure the performance of each expert. However, the choice of most appropriate expert remains difficult in several contexts in which gold standards do not exist. To overcome this issue, we propose the analogy of the 'similarity jury,' seen as a panel of experts having to reach a decision about a complex case, i.e. ranking the semantic similarity of a set of concepts. In this jury, experts are not human beings, but computable measures of similarity. Formally, the similarity function sim quantifies the semantic similarity of a pair of geographic concepts c_a and c_b ($sim(c_a, c_b) \in [0, 1]$). Set P contains all concept pairs whose similarity needs to be assessed, while set S contains all the existing semantic similarity measures.

Function sim enables the ranking of a set P of concept pairs, from the most similar (e.g. *mountain* and *peak*) to the least similar (*mountain* and *wetland*). These rankings $rank_{sim}(P)$ are used to assess the cognitive plausibility of sim against the human-generated ranks $rank_{hum}(P)$. The cognitive plausibility of sim is therefore the Spearman's correlation $\rho \in [-1, 1]$ between $rank_{hum}(P)$ and $rank_{sim}(P)$. If ρ is close to 1 or -1, sim is highly plausible, while if ρ is close to 0, sim shows no correlation with human behaviour.

A similarity jury J is defined as a set of lexical similarity measures $J = \{sim_1, sim_2 \ \dots \ sim_n\}$, where all $sim \in S$. For example, considering the eight measures in Table 2.2, jury a has two members ($J_a = \{jcn, lesk\}$), while jury b has three members ($J_b = \{jcn, res, wup\}$). Several techniques have been discussed to aggregate rankings, using either unsupervised or supervised methods [48]. However, Clemen and Winkler [48] stated that simple mathematical methods, such as the average, tend to perform quite well to combine expert judgements in risk assessment. Thus for this initial exploration, we define the rankings of jury J as the *mean of the rankings* computed by each of its individual measures $sim \in J$. For example, if three measures rank five concept pairs as $\{1, 2, 3, 4, 5\}$, $\{2, 1, 4, 3, 5\}$ and $\{1, 2, 5, 3, 4\}$, the means are $\{1.3, 1.7, 4, 3.3, 4.7\}$, resulting in the new ranking $\{1, 2, 4, 3, 5\}$.

Furthermore, we define ρ_{sim} as the correlation of an individual measure sim (i.e. a jury of size 1), and ρ_J the correlation of the judgement obtained from a jury J. If $\forall sim \in J : \rho_J > \rho_{sim}$, the jury has *succeeded* in giving a more cognitively plausible similarity judgement. On the other hand, when $\exists sim \in J : \rho_J < \rho_{sim}$, the jury has *failed*, being less plausible than its constituent measure sim. A jury J enjoys a *partial success* against sim if $\rho_J > \rho_{sim}$, where $sim \in J$. Similarly, a jury obtains a *total success* if it outperforms all of its members, $\forall sim \in J : \rho_J > \rho_{sim}$.

5.7.2 Experiment setup

In this section we describe the evaluation the similarity jury, by comparing 154 juries with eight individual measures, through an experiment on lexical similarity of OSM concepts. In order to study the similarity jury, we selected an existing dataset of human-generated similarity rankings on 54 pairs of geographic concepts, collected by Rodríguez and Egenhofer [250] from 72 human subjects. This dataset represents a high-quality sample of human judge-

ments on geospatial similarity, covering large natural entities (e.g. 'mountain,' 'forest') and man-made features (e.g. 'bridge,' 'house'). The concepts of the human-generated dataset were manually mapped onto the corresponding concepts in OSM, based on their lexical definitions.

To explore the performance of a similarity jury versus individual measures, we have selected a set of eight sim term-to-term WordNet-based measures, $S = \{jcn, lch, lesk, lin, path, res, vector, wup\}$ (see Table 2.2). The open source project WordNet::Similarity[12] implements all of these measures, and was used to compute the similarity scores [234]. As the focus in this study is on the comparison of short segments of text, rather than individual words, the word similarity scores are combined using the technique developed by Corley and Mihalcea [54]. Since the OSM Wiki website holds about 1,900 concept definitions, the complete, symmetric similarity matrix for OSM concepts would contain about 1.8 million rankings.

In the context of risk assessment, large panels with more than five experts do not seem to outperform smaller ones [76]. Therefore, we consider the range of jury sizes $|J| \in [2, 4]$ to be appropriate for this study. All the subsets of S of cardinality two, three, and four were computed, resulting respectively in 28, 56, and 70 juries, for a total of 154 juries. The experiment was carried out through the following steps:

1. Compute $rank_{sim}(P)$ for the eight measures on the OSM definitions.
2. Combine the individual measures into 154 jury $rank_J(P)$ by averaging the $rank_{sim}(P)$ of their members.
3. Compute cognitive plausibility against human-generated rankings for the eight individual measures (ρ_{sim}) and the 154 juries (ρ_J).
4. Compute partial and total success ratio for juries containing a given sim.

The next section discusses the results obtained from this experiment, drawing conclusions about the lexical similarity jury.

5.7.3 Experiment results

Table 5.11 summarises the results of this experiment, showing the success and total success ratio of the juries containing a given sim, and the total success for each measure. The table shows the cognitive plausibility ρ for each measure sim, computed against the human rankings. It is possible to note that measures *jcn*, *res*, and *lch* have the highest cognitive plausibility. The jury results are grouped by jury cardinality (2, 3, and 4), and overall results (all). The results of the experiment are also displayed in Figure 5.8, which shows the success ratio of the juries grouped by their cardinality. For example, 80.1% of all juries of cardinality 3 containing the measure *jcn* are better than *jcn* in isolation. These results show a clear pattern: most juries enjoy a partial success over a given

[12] `http://wn-similarity.sourceforge.net`

		Jury containing sim (%)								
	$\|J\|$	jcn	lch	lesk	lin	path	res	vector	wup	mean
Partial success $\rho_J > \rho_{sim}$	2	69.3	62.9	84.6	55.0	60.4	79.6	55.0	66.4	66.6
	3	80.1	68.5	84.4	60.5	58.8	86.5	61.8	72.6	71.6
	4	84.4	73.1	83.7	60.4	61.9	87.2	65.6	73.9	73.8
	all	81.3	70.4	84.0	59.8	60.7	86.2	63.2	72.7	72.3
Total success $\forall sim \in J : \rho_J > \rho_{sim}$	2	46.1	42.5	35.7	43.9	42.1	34.6	35.0	42.1	40.2
	3	43.9	37.3	34.9	40.4	31.0	32.0	33.1	36.4	36.1
	4	39.7	32.7	33.9	35.0	28.6	29.5	30.9	33.1	32.9
	all	41.8	35.3	34.4	37.8	30.9	30.9	32.1	35.2	34.8
Plausibility	ρ	.72	.68	.45	.56	.66	.69	.56	.64	.62

Table 5.11: Results of the evaluation of the lexical similarity on 154 juries. For example, juries of cardinality 2 containing *jcn* obtain a partial success in the 69.3% of the cases.

sim ($> 59.8\%$), while a minority of the juries obtain total success on all of their members ($< 41.8\%$). It is interesting to note that, in the experimental results, the plausibility of a jury is never inferior to that of all of its members, $\exists sim \in J : \rho_J < \rho_{sim}$.

The jury size has a clear impact on the success rate. Small juries of cardinality 2 tend to have a lower partial success ($mean = 66.6\%$), than those with 3 and 4 members (respectively 71.6% and 73.8%). Therefore larger juries have higher chances to obtain partial success over an individual measure. On the other hand, an opposite trend can be observed in the total success of a jury over all of its member measures. Juries of cardinality 2 tend to have a higher total success rate ($mean = 40.2\%$), compared with larger juries ($mean = 36.1\%$ for cardinality 3, and 32.9% for cardinality 4). As larger juries include more measures, it is more likely that one member outperforms the jury.

This empirical evidence shows that in 93.2% of the cases, the jury performs better than the average of the cognitive plausibility of its members, which would be by definition always lower than the plausibility of the best member: if the jury were simply returning the mean plausibility, its total success rate would always be 0%. By averaging the rankings, the jury reduces the weight of individual bias, converging towards a shared judgement. Such shared judgement is not necessarily the best fit in absolute terms, but tends to be more reliable than most individual judgements.

Given that we are measuring the cognitive plausibility of these similarity measures by the correlation with human rankings, the relationship between ρ of sim and the jury success ratio needs to be discussed. Interestingly, the cognitive plausibility ρ_{sim} shows no correlation with the jury partial and total success ratios (Spearman's $\rho \approx .1$). This suggests that even measures with high plausibility (such as *jcn* and *res*) still benefit from being combined with other measures. For example, the most plausible measure is *jcn* ($\rho = .72$), so it would be reasonable to expect a low success ratio, given that the measure is the most qualified expert in the panel. This expectation is not met: *jcn* shows a high partial and total success ratio (respectively 81.3% and 41.8%). The juries not only outperform individual measures in most cases, but can also obtain higher cognitive plausibility than its best member.

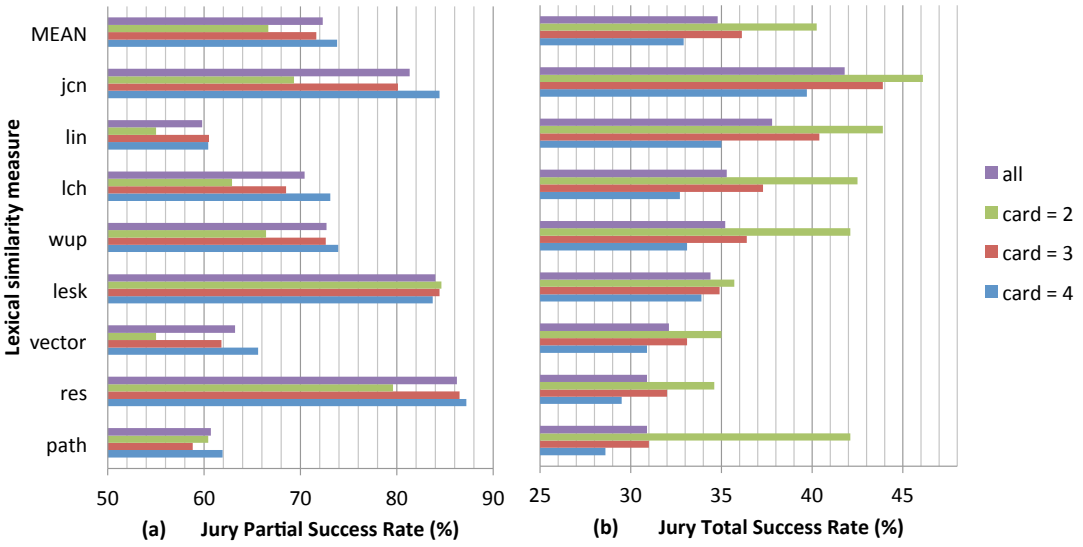

Figure 5.8: Results of the lexical jury experiment: (a) partial success of the jury versus an individual measure; (b) total success of the jury versus all its member measures. *MEAN*: mean of success rates; *card*: cardinality of jury J.

In this section we have proposed the analogy of the 'similarity jury,' a combination of semantic similarity measures. The idea of jury was then evaluated in the context of lexical similarity for OSM geographic concepts, using eight WordNet-based semantic similarity measures. Based on empirical results, the following conclusions can be drawn:

- In the context of the lexical similarity of geographic concepts, a similarity jury J is generally more cognitively plausible than its individual measures sim (partial success ratio $> 84.6\%$).
- A jury J is generally less cognitively plausible than the best of its members, i.e. $max(\rho_{sim}) > \rho_J$ (total success ratio $< 46.1\%$).
- In a context of limited information in which the optimal measure sim is not known, it is reasonable to rely on a jury J rather than on an arbitrary measure. The jury often outperforms even the most plausible measures.
- The similarity jury is consistent with the fact that, as Cooke and Goossens [53] pointed out, "a group of experts tends to perform better than the average solitary expert, but the best individual in the group often outperforms the group as a whole" (p. 644).

In this initial study we have investigated the general behaviour of the similarity jury, by combining term-to-term WordNet-based similarity measures sim, in the context of geographic concepts of OSM. Our findings are consistent with those in the area of expert judgement combination for risk assessment [48, 53]. This indicates that the analogy of the jury is sound in the

context of semantic similarity measures. However, in order to generalise these results, more work would be needed.

We have adopted a simple technique to combine rankings, i.e. a simple mean. More sophisticated techniques to combine rankings could be explored [244]. Furthermore, the empirical evidence presented in this thesis was collected in a specific context, i.e. the lexical similarity of the geographic concepts defined in OSM. General-purpose similarity datasets, such as that by Finkelstein et al. [78], could be used to conduct experiments across other semantic domains (see Section 2.5.5 for a survey of existing gold standards).

The importance of semantic similarity measures in information retrieval, natural language processing, and data mining can hardly be underestimated [261]. A scientific contribution can be given not only by devising new similarity measures, but also by identifying effective ways to combine existing measures. In this sense, we believe that the similarity jury represents a promising direction worth investigating further, given its potential to enhance the cognitive plausibility of computational measures of semantic similarity.

5.8 Summary

In this chapter, we reported the preliminary empirical evaluation that we carried out to assess our contribution to the semantics of grassroots project OpenStreetMap (OSM). First, we covered a preliminary assessment of our semantic enrichment of OSM, based on knowledge extracted from open geo-knowledge bases (Section 5.2). We then moved on to provide a detailed discussion of the concept of cognitive plausibility, on which our evaluation to semantic similarity relies (Section 5.3).

Using an existing similarity dataset, the MDSM evaluation dataset (Section 5.4), we performed a pilot evaluation on the two components of our Network-Lexical Similarity Measure (NetLexSiM), co-citation network similarity (Section 5.5), and the similarity of crowdsourced lexical definitions (Section 5.6). As part of the pilot evaluation on lexical similarity, we presented the concept of the 'similarity jury' (Section 5.7).

A jury is a combination of similarity measures, and has strong analogies with a panel of experts evaluating a complex, uncertain technical problem. On average, the jury performs better than individual measures. Overall, the cognitive plausibility of such approaches is comparatively high, with the upper bound $\rho = .85$ for network similarity, and $.81$ for lexical similarity, where 0 means no correlation with human-generated data, and 1 perfect correlation.

During this evaluation, the need for a more specific similarity gold standard emerged. While the MDSM evaluation dataset offers a set of human-generated rankings, it has a limited coverage (29 concepts). Furthermore, as pointed out in Section 5.4, the MDSM evaluation dataset has an issue of non-independence, stressing the need for an empirical evaluation on different data to cross-validate these preliminary results. Ideally, human similarity judgements should be expressed directly on the OSM concepts, on a larger con-

cept sample. With these objectives in mind, we designed, collected, and validated a novel geo-similarity dataset, Geo Relatedness and Similarity Dataset (GeReSiD), described in the next chapter.

CHAPTER

SIX

EVALUATION

6.1 Overview

In Volunteered Geographic Information (VGI), semantics plays a crucial role in enabling the usage of crowdsourced geographic data. Concentrating our efforts on OpenStreetMap (OSM), we have developed a novel semantic resource, the OSM Semantic Network. This network was then used as a semantic ground to devise a measure to compute the semantic similarity of concepts, the Network-Lexical Similarity Measure (NetLexSiM). To assess the merits and limitations of our contribution, we observe the cognitive plausibility, i.e. the correlation with human judgement, as an indication of the effectiveness of semantic similarity measures.

The pilot experiments conducted on NetLexSiM in Chapter 5 highlighted the need for a more focused evaluation. While the MDSM evaluation dataset offers a good starting point, it has a rather small coverage (33 concepts), and the manual mapping with OSM leaves out four concepts. A more salient indication of cognitive similarity should rely on human judgements expressed directly on the OSM tags, widening the coverage to a higher number of sample concepts. Hence, this chapter describes a novel geo-similarity dataset, the Geo Relatedness and Similarity Dataset (GeReSiD), and an extensive evaluation of NetLexSiM, in its network and lexical components, and then combines them into a hybrid measure. This empirical evidence provides practical advice to compute the semantic similarity or relatedness of OSM geographic concepts.

This chapter is organised as follows. Having obtained promising results in the pilot evaluation, we design, collect and validate a new gold standard, the GeReSiD (Section 6.2). This dataset provides human-generated judgements on OSM geographic concepts, highlighting the difference between semantic similarity and relatedness. GeReSiD offers more solid ground for the evaluation of co-citation network similarity algorithms (Section 6.3), and lexical similarity (Section 6.4). The results of this evaluation are combined with the pilot evaluation through a meta-analytic method, showing general trends in the data.

The two components to similarity, network and lexical, are then combined into a hybrid similarity measure, which consistently outperforms its individual components (Section 6.5). These aggregated results indicate the high cognitive plausibility of our approach to similarity and relatedness for volunteered geographic concepts. Finally, we depict a preliminary evaluation of our

approach to holistic viewport similarity, pointing out a promising and alternative direction for geo-semantic similarity research (Section 6.6).

6.2 Geo Relatedness and Similarity Dataset (GeReSiD)

This section presents a dataset of human judgements on the semantic similarity and relatedness of OSM geographic concepts, called Geo Relatedness and Similarity Dataset (GeReSiD), collected to evaluate our measure NetLexSiM. In order to evaluate a semantic similarity measure, the two main approaches consist of (1) cognitive plausibility assessment, or (2) task-based evaluation. As discussed in Section 5.3, we opt for the cognitive plausibility approach, relying on direct similarity judgement. Originally developed in linguistics, this approach has been widely used to study semantic similarity [252, 245, 78, 250]. Typically, human subjects are asked to order or rate pairs of concepts, ranging from very dissimilar to very similar. Such similarity datasets are then used as gold standards to compute the cognitive plausibility of a given measure (see Section 2.5.5 for a detailed survey of existing similarity datasets).

The cognitive plausibility approach offers a general method to establish a 'psychological ground truth', against which computational models can be compared [172]. Caution is needed, because semantic judgements are well known for being highly subjective, and this variability arises in the psychological assessment [214]. However, the issue of subjectivity is also present in application-based, indirect evaluations: all studies involving semantics rely on some human judgement on the concepts being analysed, and perfect agreement is very unlikely to be reached. For this reason, we deem that a psychological test is appropriate for the context of this study on semantic similarity and relatedness. The objectives of this psychological evaluation are the following:

- Obtain human judgements on the semantic similarity and relatedness between geographic concepts, commonly found in OSM maps.
- Capture explicitly the difference between semantic similarity and semantic relatedness (see Section 3.5).
- Cover a sample of geographic concepts larger than existing similarity datasets, including natural and man-made concepts (see Section 2.5.5).
- Include a sample of evenly distributed similarity/relatedness judgements, ranging from near-synonymity to no relationship between the concepts.

These objectives guide our attempt to design a novel gold standard for semantic relatedness and similarity, tailored to the concepts utilised in the OSM vector dataset and defined on the OSM Wiki website. Section 6.2.1 covers the design of the online survey, while Section 6.2.2 outlines the results obtained from the online survey.

6.2.1 Survey design

The Geo Relatedness and Similarity Dataset (GeReSiD) was collected through an online survey, in an uncontrolled environment. This section reports on the process through which this semantic similarity and relatedness online survey was designed and executed.

Online surveys are a powerful research tool, with well-known advantages and disadvantages [29, 311]. Given the focus of this study on volunteered geospatial data, subjects involved in VGI projects such as OSM represent an ideal virtual community of map users and producers to conduct a psychological evaluation. Effective semantic measures for OSM should be able to match the opinion of VGI active members. In this sense, an online survey is an inexpensive and effective way to reach this community.

A cross-disciplinary consensus exists on the fact that semantic judgements are affected by the *context* in which the concepts are considered [250, 134]. Rodríguez and Egenhofer [250] asked their subjects to rank geographic concepts in the following contexts: 'null context', 'play a sport', 'compare constructions', and 'compare transportation systems'. The subjects' attention is therefore focused on specific aspects of the concepts being analysed, rather than on the concepts in an unspecified setting.

Although context affects the assessment of semantic similarity, in this survey we aim at capturing the overall difference between semantic *similarity* and *relatedness* of concepts, without focusing on specific aspects of the conceptualisation. This comparison is an important research topic, frequently mentioned but never addressed directly through empirical evaluation (see Section 3.5). Introducing specific contexts into our survey would increase the complexity of the survey, making the comparison between similarity and relatedness problematic. For example, adding a specific context does not increase the inter-subject agreement: in their evaluation, Rodríguez and Egenhofer [250] report a considerably lower association between subjects in the case of context-specific questions (mean Kendall's W being .5), than with a-contextual questions (mean W = .68). Moreover, specific contexts would introduce biases.

As a solution to these issues, we frame the evaluation in the general context of popular *web maps*, in which geographic concepts are most frequently visualised and utilised by users. This way, the subjects are induced to use their own conceptualisation of the geographic entities. Subjects are free to choose what properties they consider most relevant to the comparison, and the mean of their ratings quantifies the perceived inter-subject similarity and relatedness of the concepts. While the study of the context is beyond the scope of this survey, it certainly represents an interesting direction for future work.

As the concepts included in this survey are taken from OSM, a suitable sample of concepts had to be selected. We defined a sample set P of pairs of geographic concepts, i.e. OSM tags (key, values, and proposed tags). To date, the OSM Semantic Network contains 1,503 keys, 2,047 values, and 784 proposed tags, for a total of 4,334 concepts. To be included, a tag had to be clearly intelligible, defined on the OSM Wiki website, as culturally-unspecific as pos-

sible, and present in the actual OSM vector map. Following these criteria, we included in P 400 tags, with a wide range of natural and man-made entities, such as 'sea', 'lighthouse', 'landfill', 'valley', and 'glacier'.

In order to detect issues in the survey, a pilot study was then conducted on 12 graduate students at University College Dublin, showing 100 pairs of randomly selected pairs P_{rand} on a computer in a controlled environment. Each pair was associated with a 5-point Likert scale, ranging from low to high similarity. The subjects were then interviewed informally, to obtain direct feedback about the survey. Several useful observations were obtained from this pilot survey. First, most subjects found the test too long. A smaller sample size had to be selected, considering a trade-off between number of pairs and the completion time, in order to ensure that enough subjects would complete the task without losing attention. Based on the opinion of subjects, we identified 50 pairs as the maximum size of the task, with a completion time of around five minutes, suitable for an unpaid online questionnaire.

In the OSM semantic model, tags are made of a key and a value (e.g. *amenity=school*). In the pilot survey, this formalism had to be explained to the subjects, who generally found it confusing. For example, the psychological comparison between *amenity=school* and *amenity=community_centre* was influenced by the shared word 'amenity', which is highly generic and ambiguous. To make the dataset independent from the peculiar OSM tag structure, we defined short labels for all the 400 concepts. For example, *amenity=food_court* was labelled as 'food court', *shop=music* as 'music shop'.

The fully random set of 50 pairs P_{rand} used in the pilot survey obtained a distribution heavily skewed towards low similarity and relatedness. To reach a more uniform distribution, we introduced a manual selection in the process. In order to obtain an even distribution in the resulting relatedness and similarity scores, we have manually extracted from the pilot survey a set of 50 pairs rated by the 12 subjects as highly similar/related pairs (P_{high}), and 50 middle similarity/relatedness pairs (P_{med}).[1] The final set of 50 pairs for the questionnaire P_q was assembled from the following elements:

- 16 high-similarity/relatedness pairs (random sample from P_{high})
- 18 middle-similarity/relatedness pairs (random sample from P_{med})
- 16 low-similarity/relatedness pairs (random sample from P)

The pilot survey showed clearly that assigning both the relatedness *and* the similarity tasks to the same subject was impractical, and was deemed highly confusing by all subjects. Thus, in order to collect reliable judgements on similarity and relatedness, we defined two separate questionnaires, one on relatedness (Q_{REL}), and one on similarity (Q_{SIM}). The two questionnaires

[1] It is worth noting that while the selection of highly similar/related pairs is intuitive, middle-similarity/relatedness pairs is more challenging, and requires dealing with highly subjective conceptualisation. This aspect is reflected in the survey results (see Section 6.2.2).

were identical, with the exception of the description of the task, and the labels on the Likert scale (one with a 'dissimilar-similar' scale, with other with 'unrelated-related').

To avoid terminological confusion, the survey was named 'Survey on comparison of geographic concepts', without mentioning either 'similarity' or 'relatedness' in the introductory text. The examples used to illustrate semantic relatedness (*apples - bananas*, *doctor - hospital*, *tree - shade*) and similarity (*apples - bananas*, *doctor - surgeon*, *car - motorcycle*) were based on those by Mohammad and Hirst [210]. A random redirection to either Q_{REL} or Q_{SIM} was then implemented to ensure the random sampling of subjects into two groups, one for similarity and one for relatedness. As the similarity judgement was reported as more difficult than relatedness, we have set the probability of a random redirection to Q_{SIM} at $p = .65$, to obtain more responders for similarity. Each subject has only been exposed to one of the two questionnaires.

Six general demographic questions about the subject were included: age group, mother tongue, gender, and continent of origin. A textbox was available to type feedback and comments about the survey. The core of each questionnaire was the seventh question, i.e. the relatedness or similarity rating task. The subject had to rate 50 pairs of geographic concepts based on their relatedness or similarity, on a 1 to 5 Likert scale. Although the impact of size of the Likert scale, typical options being 5, 7 or 10, is debated in psychometrics, it has little impact on the rating means [58]. If the concepts were not clear to the user, a 'no answer' option had to be selected. The full text of the online questionnaires is reported in Appendix A.2.

Another aspect discussed in the similarity psychological literature is the counterintuitive fact that similarity judgements tend to be asymmetric (e.g. $s(building, hospital) \neq s(hospital, building)$) [295]. As this aspect is outside the scope of this study, the order in each pair (a, b) was randomised to limit the symmetric bias, i.e. the potential difference between $s(a, b)$ and $s(b, a)$ from the subject's perspective. Moreover, a fixed presentation order of pairs can trigger specific semantic associations between concepts, and would reduce the quality of the last pairs, rated when the subjects are more likely to be tired. To reduce this sequential-ordering bias, the presentation order of the pairs was randomised automatically for each subject at the Web interface level.

At the end of the design process, the survey dataset contained 50 pairs of geographic concepts to be rated on 5-point Likert scales, including 97 OSM concepts, with three concepts being repeated twice. The pairs were selected to ensure an even distribution between low, medium and high similarity/relatedness. The rating was to be executed in two independent questionnaires, one for semantic similarity (Q_{SIM}) and one for semantic relatedness (Q_{REL}), randomly assigned to the human subjects. In February 2012, the survey was implemented on the LimeService website,[2] and disseminated in OSM and Geographic Information Systems (GIS) forums and mailing lists.

[2] `http://www.limeservice.com`

6.2.2 Survey results

This section describes the analysis of the survey, and the resulting dataset, the Geo Relatedness and Similarity Dataset (GeReSiD). The online questionnaires on relatedness and similarity received 305 responses, 124 for relatedness and 181 for similarity.

Given the nature of online surveys, particular attention has to be paid to the agreement between the human subjects, and the detection of unreliable and random answers. In this survey, raters expressed quantitative judgements on semantic relatedness and similarity on a 5-point Likert scale. Two important psychometric aspects to be discussed are the interrater reliability (IRR) and the interrater agreement (IRA) [171]. IRR considers the *relative* similarity in ratings provided by multiple raters (i.e. the human subjects) over multiple targets (i.e. the concept pairs), focusing on the order of the targets. IRA, on the other hand, captures the *absolute* homogeneity between the ratings, looking at the specific rating chosen by raters.

Several indices have been devised to capture IRR and IRA in psychometric surveys [27, 171]. Most indices range between 0 (total disagreement) and 1 (perfect agreement). For example, the ratings of two raters on three targets $\{1, 2, 3\}$ and $\{2, 3, 4\}$ obtain a $IRR = 1$ and $IRA = 0$: the subjects agree perfectly on the ordering of the targets, while disagreeing on all absolute ratings. LeBreton and Senter [171] recommend using several indices for IRR and IRA, to avoid the bias of any single index. We thus include the following indices of IRA and IRR: the mean Pearson's correlation coefficient [249]; Kendall's W [149]; Robinson's A [248]; Finn coefficient [79]; James, Demaree and Wolf's $r_{WG(J)}$ [131].

The 305 responders included both native (208) and non-native English speakers (97). We observed a substantially lower inter-subject agreement when including non-native speakers ($r_{WG(J)} < .5$): the wider variability in these results is due the varying knowledge of English of these subjects, who might have associated concepts to false friends in their native language, e.g. Italian speakers may confuse the meaning of 'factory' with 'farm' ('fattoria' in Italian). Hence, they were excluded from the dataset.

Three subjects did not complete the task, and their responses were discarded. In order to detect random answers, we computed the correlation between every individual subject and the means. This way, two subjects in the similarity test showed no correlation at all with the mean ratings (Spearman's $\rho \in [-.05, .05]$), and were removed from the dataset.

The resulting dataset is summarised in Table 6.1. The table includes demographic information (age group, gender, etc), and the indices of IRR and IRA. Following Resnik [245], we consider upper bound for the cognitive plausibility of a computable measure to be the highest correlation obtained by a human rater with the means (e.g. $\rho = .92$ for relatedness). The table shows these upper bounds both for Spearman's ρ and Kendall's τ. All the IRR and IRA indices indicate very similar results, falling in the range $[.61, .67]$. Given the highly subjective nature of semantic conceptualisations, this correlation is

		Relatedness Q_{REL}	Similarity Q_{SIM}	Overall –
Responders	total N	81	122	203
Gender	Male	72	93	165
	Female	9	29	38
Age	18-25	28	39	67
	26-35	14	41	55
	36-45	12	23	35
	46-55	15	10	25
	56-65	7	9	16
	>65	5	-	5
Continent	Africa	-	3	3
	Asia	-	1	1
	Europe	58	95	153
	North America	11	20	31
	South America	-	-	-
	Oceania	12	3	15
Web map expertise	Never used	6	14	20
	Occasional user	18	33	51
	Frequent user	37	39	76
	Expert	20	36	56
IRR	mean Pearson's r	.64*	.65*	-
IRA	Kendall's W	.65*	.64*	-
	Robinson's A	.62	.61	-
	Finn coefficient	.65*	.66*	-
	$r_{WG(J)}$	.66	.67	-
Upper bound	Spearman's ρ	.92*	.93*	-
	Kendall's τ	.79*	.82*	-

Table 6.1: Results of online survey on geo-similarity and relatedness (GeReSiD). The table reports demographics, indices for interrater reliability (IRR) and interrater agreement (IRA). (*) $p < .001$.

satisfactory, and is comparable with analogous psychological surveys [250].

Given the set of concept pairs, and the set of human raters, we computed the relatedness/similarity scores as the *mean ratings*, normalised in the interval $[0, 1]$, where 0 means no relatedness/similarity, and 1 maximum relatedness/similarity. As we have stated in the survey objectives, the distribution of such scores should be as even as possible, to ensure that a semantic measure performs well across the board, and not only in a specific region of the semantic relatedness/similarity space.

A dimension that has not been addressed in existing similarity gold standards is that of the *pair agreement*, i.e. the consistency of ratings expressed by all subjects on a single pair (see Section 2.5.5). For this purpose, we adopt James, Demaree and Wolf's r_{WG}, a popular index to measure IRA on a single target, based on the rating variance [131]. Each pair in Q_{REL} and Q_{SIM} obtains an agreement measure $\in [0, 1]$, where 0 indicates a squared distribution (i.e. raters gave all ratings in equal proportion), and 1 is perfect agreement (i.e. all raters assigned exactly the same rating to the pair).

Figure 6.1 shows several statistical characteristics of the resulting dataset, for the 50 pairs in Q_{REL} and Q_{SIM}. Plot 6.1(a) shows the density of the final

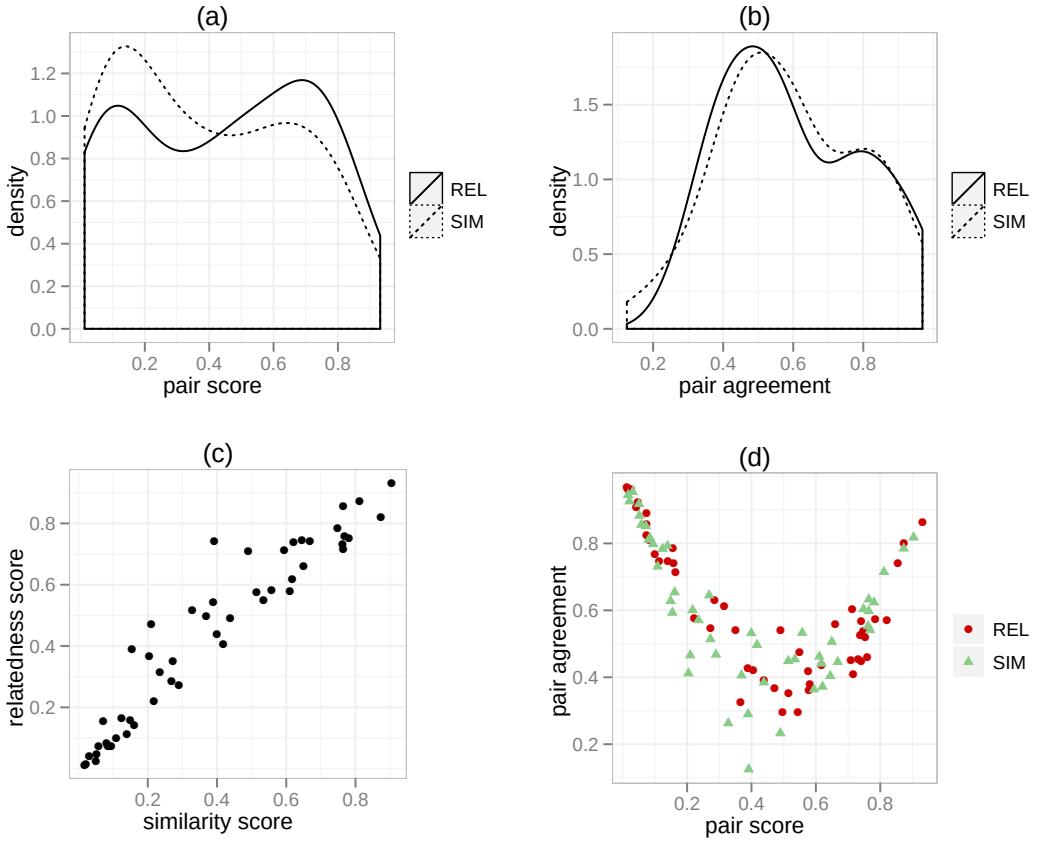

Figure 6.1: GeReSiD results. REL: semantic relatedness; SIM: semantic similarity. (a) density of pair score; (b) density of pair agreement; (c) scatterplot of relatedness versus similarity; (d) scatterplot of pair agreement and pair score.

relatedness/similarity scores, i.e. the normalised mean rankings. While the similarity is skewed towards the range $[0, .4]$, the relatedness has slightly more scores in the range $[.4, 1]$, resulting in symmetrical densities. This clearly reflects the fact that semantic similarity is a specific type of semantic relatedness, and semantic similarity is generally lower than relatedness. This can be also observed in the sum of the 50 relatedness scores ($sum = 22.01, mean = .44$) against the similarity scores ($sum = 19.5, mean = .39$). The paired Wilcoxon signed rank test [307] indicates that the relatedness scores are higher than the corresponding similarity ones, at $p < .001$. This trend is clearly visible in plot 6.1(c). Overall, these densities show that all the score range $[0, 1]$ is satisfactorily covered, i.e. the dataset does not show large gaps.

Plots 6.1(b), (c) and (d) show the properties of pair agreement (index r_{WG}), reporting the relationship between relatedness and similarity, the density of pair agreement, and the relationship between pair agreement and relatedness/similarity scores. In terms of pair agreement, relatedness and similarity follow very close patterns, with a peak $\approx .5$. This agreement might seem low, but it is largely expected, due to the subjective interpretation of the values on the Likert scale.

An explanation of this trend in the pair agreement lies in the fact that hu-

mans give consistently different ratings to the same objects: some subjects tend to be strict, and some lenient, resulting in different relative ratings, and therefore low absolute pair agreement [171]. In this regard, a clear pattern emerges from plot 6.1(d). Pair agreement tends to be high ($> .7$) at the extremes of the scores, when the relatedness/similarity judgement is very low ($[0, .25)$, no relation) or very high ($(.75, 1]$, strong relation). On the other hand, pairs with middle scores (in the interval $[.25, .75]$) tend to have low pair agreement. Relatedness and similarity do not show important differences with respect to pair agreement ($sum = 30, mean = .6$ for relatedness, $sum = 29.6, mean = .59$ for similarity). To conclude this discussion on the GeReSiD, the following points are of particular importance:

- Geo Relatedness and Similarity Dataset (GeReSiD) contains human judgements about 50 pairs, on semantic relatedness and similarity. The judgements were collected from two separate groups of 81 and 122 English native speakers, through a randomised online survey. The dataset is freely available online, released under an Open Knowledge license.[3]
- The human judgements have interrater agreement (IRA) and interrater reliability (IRR) in the interval $[.61, .67]$. Considering the type of psychological test, this is a fair agreement, indicating that the dataset can be used to evaluate computational measures of semantic relatedness and similarity for geographic concepts.
- Human subjects strongly agree on cases of very high and low semantic relationships, and tend to have lower agreement on the intermediate cases.
- Semantic relatedness and similarity are strongly correlated ($\tau = .84, \rho = .95$). Furthermore, semantic relatedness scores consistently higher than semantic similarity, confirming the more specific nature of semantic similarity.
- GeReSiD can be used as a gold standard to evaluate measures of semantic similarity and relatedness. Furthermore, it permits the empirical determination of whether a given measure better approximates similarity or relatedness. For example, NetLexSiM was designed as a similarity measure, but in some cases obtains higher cognitive plausibility for relatedness.

Having designed, collected, and validated this new dataset, tailored for OSM geographic concepts, we move on to conduct a detailed evaluation of NetLexSiM. Sections 6.3 and 6.4 proceed to describe this evaluation, relying on GeReSiD as a gold standard for semantic relatedness and similarity for OSM concepts.

[3] `http://github.com/ucd-spatial/Datasets`

6.3 NetLexSiM: evaluation of network similarity

This section reports on the evaluation we conducted to assess the network component of NetLexSiM, outlined in Section 3.6, using our GeReSiD as a gold standard. In order to evaluate the cognitive plausibility of co-citation measures applied to the OSM Semantic Network, an experiment was set up, closely following the approach used in the preliminary evaluation conducted on the MDSM evaluation dataset (see Section 5.5). The scores generated by the co-citation algorithms were compared with the similarity and relatedness scores of the 50 pairs contained in GeReSiD. The hypotheses tested through this evaluation are the following:

1. Co-citation measures applied on the OSM Semantic Network can reach high cognitive plausibility for semantic similarity and, to a more minor extent, for relatedness.

2. The similarity judgements contained in GeReSiD are consistent with those in the MDSM evaluation dataset.

Section 6.3.1 outlines the experiment setup, while Section 6.3.2 discusses the empirical results. The results are then compared with those obtained the in the preliminary evaluation through a meta-analysis in Section 6.3.3.

6.3.1 Experiment setup

Co-citation algorithms are based on the idea that, in a network, similar vertices tend to be linked from the same vertices, and tend to link the same vertices. The recursive co-citation algorithm P-Rank includes a number of co-citation algorithms [318], including among others Coupling [152], Amsler [9], and SimRank [136].

As discussed in Section 3.6, decay factor $C \in (0, 1)$ controls the propagation of similarity across the graph edges. The similarity scores, transferred from a pair of vertices to their neighbours, are multipled to C. Furthermore, $\lambda \in [0, 1]$ determines the balance between incoming and outgoing links in the similarity computation. When $\lambda = 1$, the computation of similarity considers exclusively incoming edges. By contrast, when $\lambda = 0$, the computation of similarity is based only on outgoing edges. Constant K is the number of iterations through which the iterative similarity scores are computed. When $K = 1$, P-Rank is not iterative and corresponds to classic co-citation measures. This experiment follows closely the preliminary evaluation reported in Section 5.5. To explore the performance of the co-citation algorithms, the following of the P-Rank parameters were selected:

- λ (P-Rank in-out link balance): 11 discrete equidistant levels $\in [0, 1]$.
- C (P-Rank decay constant): 9 discrete equidistant levels $\in [.1, .9]$. $C = .95$ was also included, being the optimal value in the preliminary evaluation in Section 5.5.

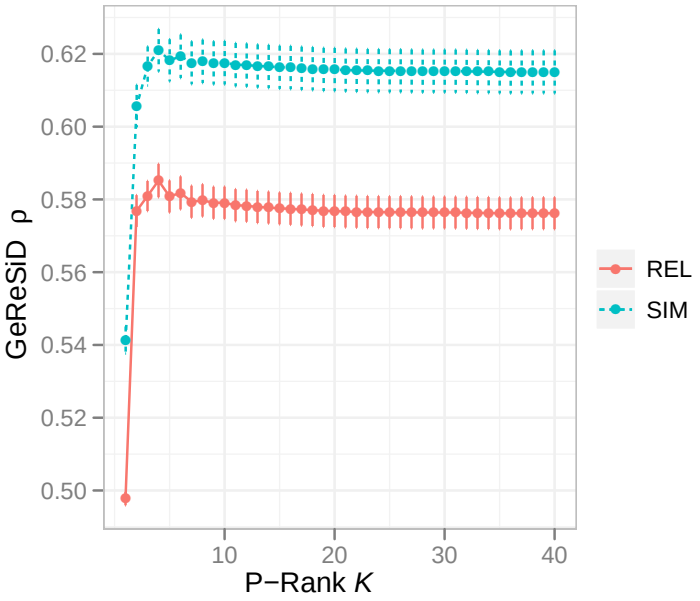

Figure 6.2: Experiment results grouped by P-Rank K, with standard error bars. All Spearman's ρ are computed on 50 pairs without ties ($p < .001$) as a measure of correlation with human behaviour.

– K (P-Rank iterations): 40 P-Rank iterations.

These parameters resulted in 4,400 unique combinations of λ, C, and K. The similarity scores were then obtained for the 100 concept pairs in GeReSiD, applying P-Rank on the OSM Semantic Network (see Section 3.3) for these 4,400 combinations. The resulting 4,400 sets of similarity scores where subsequently compared with the relatedness and similarity scores of GeReSiD. Spearman's rank correlation coefficient ρ [276] was utilised to assess the correlation between machine and human-generated scores, on the 50 pairs of similarity scores without ties. Thus, 8,800 correlations for similarity and relatedness were obtained. The statistical significance of these correlations is high in all of the 8,800 cases ($p < .001$).

6.3.2 Experiment results

The experiment resulted in 8,800 correlations between co-citation similarity measures on the OSM Semantic Network and the scores in GeReSiD. The correlation between the human dataset (GeReSiD) and the machine generated scores (co-citation algorithms) is influenced by three parameters: K, the number of P-Rank iterations; λ, the in-out link balance; C, the P-Rank decay factor. The details of these algorithms are described in Section 3.6. In order to identify general trends in the results, the correlations are grouped by the three P-Rank parameters.

As K increases, the similarity scores are closer to the theoretical, asymptotic value of P-Rank. Figure 6.2 shows the impact of K on the correlation

K $[1,\infty)$	λ $[0,1]$	C $(0,1)$	Co-citation measure	GeReSiD SIM ρ	GeReSiD REL ρ
1	0	–	Coupling [152]	.5	.47
1	.5	–	Amsler [9]	.53	.5
1	1	–	Co-citation [271]	.61	.49
40	0	.9	rvs-SimRank [318]	.46	.45
40	0	.5	–	.5	.49
40	0	.1	–	.51	.5
40	.5	.9	P-Rank [318]	.64	.6
40	.5	.5	–	.62	.6
40	.5	.1	–	.6	.58
40	.9	.8	–	.73*	.65*
40	.9	.7	–	.72	.64
40	.9	.6	–	.73*	.62
40	.9	.5	–	.72	.64
40	1	.9	SimRank [136]	.65	.55
40	1	.5	–	.64	.53
40	1	.1	–	.64	.54

Table 6.2: Cognitive plausibility of co-citation similarity measures against GeReSiD. Each correlation test includes 50 pairs of rankings without ties. All Spearman's ρ have $p < .001$. (*) Best performance.

with GeReSiD, for semantic relatedness and similarity. It is possible to notice a quick convergence with $K \in [1, 10]$, followed by a slow decline in the interval $[11, 20]$. With $K > 20$, the correlations remain stable, around the mean $\rho = .62$ for similarity, and $\rho = .58$ for relatedness.

The constant C determines how fast the similarity decays during the iterations. When $C \to 0$, the decay is fast, while $C \to 1$ implies a slow decay. The impact of C on the cognitive plausibility of the co-citation algorithms is displayed in Figure 6.3. Low values of C ($[.1, .4]$) correspond to lowest plausibility in the experiment, both for relatedness and similarity ($\rho < .58$ for relatedness, $\rho < .65$ for similarity). The middle value ($C = .5$) results in a peak for semantic relatedness ($\rho = .58$). By contrast, the best results for similarity are obtained when $C \in [.5, .9]$, with a peak at $C = .8$ ($\rho = .62$), and a drop when $C = .95$.

The third parameter that influences the results of P-Rank is λ, the balance between in and out-links in the semantic network. When $\lambda = 0$, only the out-links are considered, while $\lambda = 1$ includes only in-links. Figure 6.4 shows the impact of λ on the cognitive plausibility of P-Rank. The performance of the algorithms improve steadily as λ moves from 0 to $.9$, with a peak at $\lambda = .9$ ($\rho = .7$ for similarity, $\rho = .63$ for relatedness). When $\lambda = 1$, the performance decreases suddenly ($\rho = .63$ for similarity, $\rho = .53$ for relatedness). Hence, focusing on the best approximations to the theoretical value of P-Rank ($K = 40$), the most plausible results against GeReSiD are located in the intervals $C \in [.5, .8], \lambda \in [.8, .9]$. In this region, the mean correlation with the human rankings is $\rho = .73$ for similarity, and $\rho = .65$ for relatedness. These results confirm our first hypothesis, i.e. co-citation measures in the OSM Semantic Network reach high cognitive plausibility for semantic similarity and a lower plausibility for relatedness.

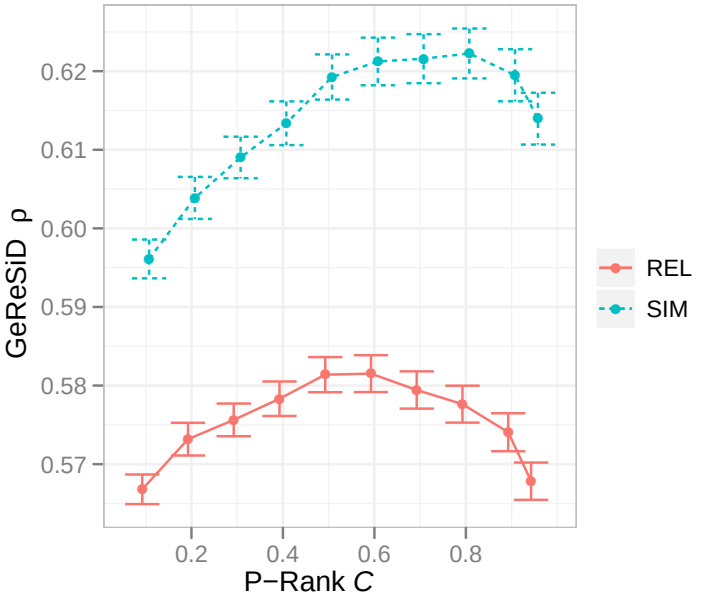

Figure 6.3: Experiment results grouped by P-Rank decay factor C, with standard error bars. Spearman's ρ is a measure of correlation with human behaviour ($p < .001$). Each correlation test includes 50 pairs of rankings without ties.

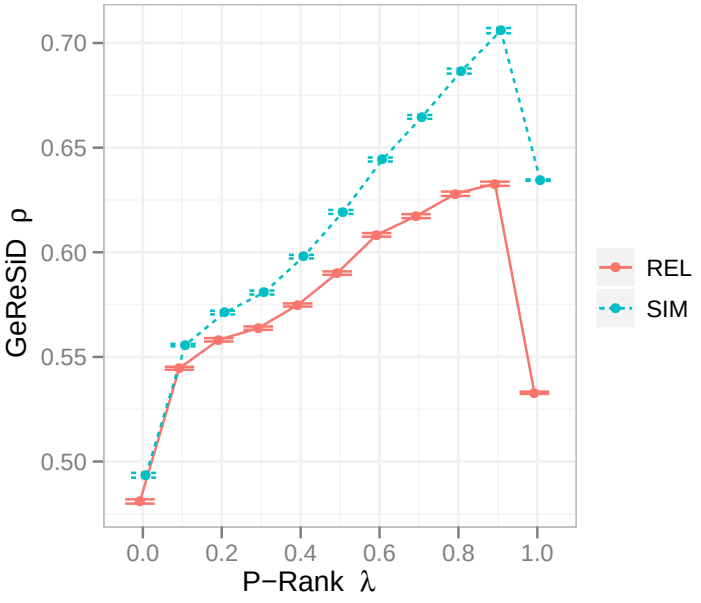

Figure 6.4: Experiment results grouped by P-Rank in-out link factor λ, with standard error bars. Spearman's ρ is a measure of correlation with human behaviour ($p < .001$). Each correlation test includes 50 pairs of rankings without ties.

		P-Rank		Row
		Non-sim	Sim	total
Humans	Non-sim	22	3	25
	Sim	3	22	25
	Column total	25	25	50

Table 6.3: Network-based similarity, contingency table of pairs judged similar (sim) or non-similar (non-sim) by the human subjects and by P-Rank on the GeReSiD dataset. Fisher's exact test: $p < .001$

In order to provide further evidence of the reliability of these results, we have conducted an analysis treating the semantic similarity judgements as categorical. The rankings generated by the human subjects and P-Rank were transformed into two categories (similar/not similar), using the average similarity ranking as a threshold. All the pairs ranked above the average rank are considered to be similar, and the others are not similar. Applying this analysis to the best case of P-Rank ($C = .8$, $\lambda = .9$, $K = 40$), we obtained the contingency table displayed in Table 6.3. Fisher's exact test applied to this 2x2 contingency table results in high statistical significance ($p < .001$). This result confirms the high cognitive plausibility of the network component of NetLexSiM.

Although optimal parameters lead to strong correlation for similarity ($\rho \approx .7$), it is beneficial to assess the cases in which the network similarity measures show a considerable discrepancy with the human-generated rankings. When $K = 40$, $C = .8$, and $\lambda = .9$, concept pair *<arts centre, bureau de change>* is ranked 33th in the set of 50 pairs by the human subjects, while the pair is ranked 6th by P-Rank. This wide gap is due to the high structural similarity of the two concepts, which are both linked to the key *amenity*, and are not densely linked to other concepts that might help the algorithm reduce their semantic similarity.

The opposite case arises with two pairs, *<city, railway station>* and *<heritage item, valley>*, which are ranked respectively 25th and 26th by the human subjects, and are ranked 44th and 45th by P-Rank. The reason for the mid-range similarity lies in the fact that railway stations are often located within cities, and some natural valleys are classified as part of the natural heritage. These weak subsumption relations are not captured by the link structure in the OSM Semantic Network, and therefore P-Rank fails to find any similarity between the pairs. In other words, the relations between these concepts are underspecified. These failures of the similarity model can be mitigated by including the lexical definitions into a hybrid computation, as it is discussed in Section 6.5.

Table 6.2 summarises the results of this evaluation, comparing the cognitive plausibility of the algorithms against GeReSiD. It is possible to note that the cognitive plausibility for similarity is consistently higher than for relatedness throughout the dataset. This can be easily explained by the fact that the co-citation algorithms are designed to measure similarity, and not relatedness. Although the cognitive plausibility for relatedness (max $\rho = .65$) may still be

appropriate for certain applications, other approaches should be devised to obtain higher plausibility. The next section compares these results with the pilot evaluation on the MDSM evaluation dataset, using a meta-analytical approach.

6.3.3 Network similarity meta-analysis

In order to draw general conclusions from the experiments on network similarity, conducted on the MDSM evaluation dataset and on GeReSiD, a meta-analysis is needed. According to Glass [93], a meta-analysis is "the analysis of analyses, ... the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). Meta-analytical approaches are very popular in the social sciences, medicine, and psychology.

In our context, a meta-analysis also has the purpose of confirming the cognitive plausibility of co-citation approaches as similarity measures applied to the OSM Semantic Network, cross validating the psychological accuracy of GeReSiD. While the cognitive plausibility in relation to the MDSM evaluation dataset is measured as the weighted mean of five sets ($\bar{\rho}$), GeReSiD's ρ is measured on a set of 50 pairs. The correlation between the cognitive plausibility for the two datasets is very strong ($\rho = .94$ on 50 pairs without ties, $p < .0001$). This confirms our second hypothesis (the similarity judgements contained in GeReSiD are consistent with those in the MDSM evaluation dataset), and validates the human judgements in the GeReSiD, showing indirectly a substantial psychological agreement with the MDSM evaluation dataset.

These similarities notwithstanding, differences between the two datasets exist. The optimal performance of P-Rank in GeReSiD is obtained with parameters $C = .8, \lambda = .9$. By contrast, the MDSM evaluation dataset is best approximated when $C = .9, \lambda = 1$, corresponding to the SimRank algorithm. The plausibility of P-Rank suddenly drops when $\lambda = 1$ in GeReSiD, which does not occur in the MDSM evaluation dataset. This difference is due to the limited information problem that affects SimRank, as Zhao et al. [318] pointed out. As SimRank only relies on in-links, vertices that have only out-links cannot obtain a similarity score. The different coverage in the two datasets can help explain these differences. P-Rank overcomes this issue by including out-links. While the MDSM evaluation dataset contains 29 concepts, GeReSiD covers 97 OSM concepts, including more concepts affected by the limited information problem.

Overall, the results we obtained with the MDSM dataset are considerably higher, with an average differential $> .1$ when compared with GeReSiD. However, given the more limited coverage of the MDSM evaluation dataset, we suggest that the results obtained with GeReSiD are to be considered as a more realistic indicator of the cognitive plausibility of co-citation algorithms. Table 6.4 summarises the results of these evaluations, showing a meta-analysis based on the Hunter-Schmidt method [129].

For each of the six co-citation similarity algorithms, the top performance

Network similarity measure	*Similarity meta-analysis* MDSM dataset $\bar{\rho}$	GeReSiD SIM ρ	Meta SIM ρ	GeReSiD REL ρ
P-Rank [318]	.76	.73*	.74 ± .01*	.65*
SimRank [136]	.85*	.65	.7 ± .08	.55
Co-citation [271]	.72	.61	.64 ± .05	.49
Amsler [9]	.67	.53	.56 ± .06	.5
rvs-SimRank [318]	.6	.51	.53 ± .04	.5
Coupling [152]	.55	.5	.51 ± .02	.47

Table 6.4: Meta-analysis of the top performance of co-citation measures, including MDSM evaluation dataset and GeReSiD. (*) Best performance.

is considered for both datasets, and then combined in a single value, with a 95% confidence interval ($p < .001$). For example, through the meta-analysis, the best performance of P-Rank with the MDSM evaluation dataset (.76) and GeReSiD (.73) are combined into .74±.01. The top performance for semantic relatedness is also included, not as a meta-analysis as there is only a single experiment on GeReSiD. From this empirical evidence, the following conclusions on co-citation algorithms in the context of the OSM Semantic Network can be drawn:

- Recursive co-citation algorithms reach high cognitive plausibility for semantic similarity, with a maximum ρ = .74 ± .01. To a lesser degree, they also approximate semantic relatedness, with an upper bound of ρ = .65.
- The best results are obtained by SimRank (.7 ± .08) and P-Rank (.74 ± .01), when similarity flows with slow decay ($C \in [.8, .9]$), and favouring incoming links, as opposed to outgoing links ($\lambda \in [.9, 1]$).
- Among the non-recursive algorithms, classic Co-citation [271] reaches the highest plausibility (.64 ± .05).
- Network-based measures specifically devised for semantic relatedness must be explored to reach higher plausibility.

The empirical evidence summarised in this meta-analysis confirms that co-citation measures are appropriate to capture the network similarity in the OSM Semantic Network, and be used as the network component of NetLexSiM. The semantic similarity of concepts is extracted purely from the link patterns between the vertices of a directed graph. The next section describes the GeReSiD-based evaluation of the other similarity component, involving the lexical definitions of concepts.

6.4 NetLexSiM: evaluation of lexical similarity

This section discusses the evaluation we have conducted on the lexical similarity component of NetLexSiM, outlined in Section 3.7. While Section 5.6

reported on the pilot evaluation of lexical measures against the MDSM evaluation dataset, this section reports on two new experiments, using GeReSiD as a gold standard. The hypotheses that this evaluation aims at validating are the following:

1. The lexical definitions of OSM concepts enables the computation of a more plausible semantic similarity measure than the OSM tags.
2. Our knowledge-based bag-of-words (BOW) approach can reach high cognitive plausibility for both similarity and relatedness.

First, WordNet-based similarity measures are applied directly on the OSM tags contained in GeReSiD, relying on a manual mapping between tags and WordNet synsets (Section 6.4.1). The limited cognitive plausibility of this approach confirms the necessity of including the concept lexical definitions into the computation. Second, we apply our lexical similarity technique on the concept lexical definitions, obtaining better results (Section 6.4.2). Finally, a meta-analysis summarises this empirical evidence, indicating a set of guidelines to compute the lexical semantic similarity for OSM concepts, the second component of NetLexSiM (Section 6.4.3).

6.4.1 Tag-based experiment

This experiment aims at investigating the cognitive plausibility of WordNet-based similarity measures when applied directly to OSM tags, testing our first hypothesis, i.e. the lexical definitions of OSM concepts enable the computation of a more plausible semantic similarity than the OSM tags. The evaluation of this approach with the MDSM evaluation dataset shows no statistically significant correlation with the human rankings (see Section 5.6). In order to evaluate the WordNet similarity measures directly on the tags, the 97 OSM concepts contained in GeReSiD were manually mapped to the corresponding WordNet synsets. For example, the OSM tag *waterway=dock* is semantically equivalent to the WordNet synset *dock#n#3*. Subsequently, ten WordNet-based measures were computed on the GeReSiD 50 pairs for relatedness, and the 50 pairs for similarity. These similarity measures were described in detail in Section 2.5.4. The correlations of these similarity scores with the GeReSiD human scores were obtained, and are reported in Table 6.5. For each WordNet similarity measure, the table shows the tie-corrected Spearman's ρ with the similarity and relatedness, the number of tied rankings, and the corresponding levels of statistical significance.

While some measures obtained relatively high plausibility (e.g. *hso*), others resulted in weak correlations, showing very low cognitive plausibility. The statistically significant results indicate ρ in the interval $[.28, .53]$. The top performing measures, both for similarity and relatedness, are *hso*, *vector*, and *vectorp*, obtaining $\rho \in [.43, .53]$. The other measures obtain a considerably lower cognitive plausibility ($\rho < .34$), indicating no clear convergence towards the human-generated dataset. For eight measures out of ten, semantic relatedness

WordNet measure	Tied ranks	GeReSiD SIM ρ	GeReSiD REL ρ
hso	23	.53†††	.47†††
vectorp	9	.44††	.47†††
vector	5	.43††	.44††
lesk	3	.33†	.37††
lch	15	.33†	.34†
path	20	.33†	.34†
lin	17	.25	.28†
wup	12	.25	.28
jcn	19	.22	.25
res	32	.18	.20

Table 6.5: Cognitive plausibility of WordNet similarity measures applied on single synsets. Each correlation test includes 50 pairs of rankings. Statistical significance: (†) $p < .05$; (††) $p < .01$; (†††) $p < .001$.

is better approximated than similarity. This experiment shows the inadequacy of WordNet-based measures applied directly to tags. This issue might be due to insufficient detail in the WordNet taxonomic structure in relation to geographic concepts. To overcome this issue, the next section describes the evaluation of the lexical component of NetLexSiM (see Section 3.7), which relies on lexical definitions of concepts, and obtains considerably higher cognitive plausibility.

6.4.2 Definition-based experiment

Given the unsatisfactory results obtained through WordNet similarity measures applied on OSM tags, we evaluate our definition-based approach to lexical similarity, the lexical component of NetLexSiM. This approach, outlined in Section 3.7, consists of extracting semantic vectors from the lexical definitions, and then comparing them using term-to-term WordNet-based measures. An overall definition-to-definition similarity measure is obtained by combining the term-to-term similarity matrix using paraphrase detection techniques. In the pilot evaluation (see Section 5.6) this approach shows promising results when compared against the MDSM evaluation dataset. In this section, we replicated that experiment on GeReSiD.

The experiment setup consists of a set of combinations $\{POS, C, sim_t, sim_v\}$. The four parameters include ten WordNet term-to-term similarity measures sim_t (see Section 2.5.4), three text corpora C (Null, OSM Wiki website, and the Irish Independent, described in Section 5.6), three part-of-speech (POS) filters (*NN*, *NN VB*, and *VB*), and two vector-to-vector similarity measures (*com* by Corley and Mihalcea [54] and *fes* by Fernando and Stevenson [75]), for a total of 180 cases.

As the distributions of ρ for the algorithm parameters tend to be heavily skewed, we adopt the median $\tilde{\rho}$ as a robust estimator of central tendency, reporting the 25% and 75% quartiles for each parameter. As already noticed in the pilot experiment, verbs used in isolation (POS = *VB*) do not show any

Param Name	Param Value	GeReSiD SIM median $\tilde{\rho}$	GeReSiD SIM 25%-75% quartiles	GeReSiD SIM max ρ	GeReSiD REL median $\tilde{\rho}$	GeReSiD REL 25%-75% quartiles	GeReSiD REL max ρ
sim_v	com	.61	.56 .64	.74	.62	.57 .65	.74
	fes	–	– –	–	–	– –	–
POS	NN	.61*	.56 .64	.74*	.64*	.57 .66	.74*
	NN VB	.61*	.56 .64	.73	.61	.56 .65	.73
	VB	–	– –	–	–	– –	–
C	Irish Indep	.64*	.62 .72	.74	.64*	.62 .71	.74*
	Null	.58	.54 .62	.64	.61	.56 .65	.68
	OSM Wiki	.58	.52 .62	.65	.6	.52 .64	.67
sim_t	vector	.64*	.64 .71	.74*	.66	.65 .7	.72
	path	.64*	.64 .71	.73	.68*	.66 .72	.74*
	lch	.62	.62 .7	.73	.66	.65 .7	.74*
	hso	.61	.61 .66	.71	.64	.63 .67	.71
	wup	.6	.57 .62	.66	.62	.59 .64	.69
	res	.6	.59 .62	.64	.62	.6 .62	.64
	lesk	.56	.56 .62	.64	.59	.57 .63	.65
	vectorp	.54	.52 .6	.64	.56	.53 .59	.63
	jcn	.5	.49 .55	.59	.51	.49 .55	.58
	lin	.48	.45 .5	.56	.46	.46 .5	.55
all	–	.61	.56 .64	.74	.62	.57 .65	.74

Table 6.6: Summary of results of the lexical similarity experiment on GeReSiD. Each correlation test includes 50 pairs of rankings without ties. The central tendency of each parameter is summarised by the median $\tilde{\rho}$, its lower and upper quartiles, and its maximum value. Statistical significance for all cases: $p < .001$. (*) Best performance.

correlation with the human dataset, resulting in $\rho \in [.01, .16]$, with $p > .1$. Similarly, the vector-to-vector sim_v *fes* obtains low correlation with GeReSiD, both for relatedness ($\tilde{\rho} = .22$) and similarity ($\tilde{\rho} = .26$). Unlike the *com* measure, several cases where sim_v is equal to *fes* are non-significant ($p > .05$). Being not sufficiently statistically significant, cases in which POS=*VB* or sim_v =*fes* were excluded from the analysis.

The results are summarised in Table 6.6, which for each parameter reports median, quartiles, and maximum ρ for semantic similarity and relatedness. Overall, our approach to comparing OSM concepts obtains consistent results both for similarity ($\tilde{\rho}_{SIM} = .61$) and relatedness ($\tilde{\rho}_{REL} = .62$), while the upper bound is $\rho = .74$ for both similarity and relatedness. Unlike the network similarity measures, the lexical similarity approximates semantic relatedness slightly better than similarity. This is consistent with the findings of the tag-based experiment, in which WordNet-based measures obtained higher cognitive plausibility with relatedness than similarity (see Section 6.4.1). These trends are visible in Figure 6.5, which depicts the distribution of ρ for semantic similarity (SIM), summarising each distribution with a boxplot. Semantic relatedness (REL) is reported in Figure 6.6. The boxplots represent the smallest ρ, the 25% quartile, the median, the 75% quartile, and largest $\bar{\rho}$. For the sake of completeness, the mean of each distribution is marked as a white diamond in both figures.

The four parameters that influence the algorithm results are

$\{POS, C, sim_t, sim_v\}$. The vector-to-vector measure sim_v determines the strategy to compute the similarity of semantic vectors. While *fes* did not show satisfactory cognitive plausibility, *com* obtained more promising results. The POS filter selects the terms to be included in the semantic vectors. Excluding the analysis of verbs in isolation (*VB*), *NN* and *NN VB* show a very close cognitive plausibility, although *NN* performs better for relatedness ($\tilde{\rho}_{REL} = .64$) than for similarity ($\tilde{\rho}_{SIM} = .61$).

The text corpus C is utilised to assign semantic weights to the terms. The cognitive plausibility obtained by the *Null* and *OSM Wiki* corpora is largely comparable, performing marginally better for relatedness ($\tilde{\rho}_{REL} \approx .6$) than for similarity ($\tilde{\rho}_{SIM} = .58$). By contrast, the corpus extracted from the Irish Independent, containing news stories, outperforms the other corpora, resulting in a higher cognitive plausibility ($\tilde{\rho} = .64$ both for similarity and relatedness).

The fourth parameter which has a high impact on the results is the WordNet-based term-to-term measure sim_t. Measures *vector*, *path*, *lch*, and *hso* fall in the top tier, with an upper bound $\rho \geq .7$, and a median $\tilde{\rho} > .6$ for both similarity and relatedness. All the other measures perform in a less satisfactory way, with an upper bound $\rho \in [.56, .66]$ for similarity and $[.55, .69]$ for relatedness, and a lower median in the interval $[.48, .6]$ for similarity and $[.46, .62]$ for relatedness. As it is possible to observe in Figures 6.5 and 6.6, after the top cluster of these four term-to-term measures, the performance drops visibly, reaching a minimum with *lin* (median $\approx .47$, upper bound $\approx .55$). The other measures (*wup*, *res*, *lesk*, *vectorp*, and *jcn*) fall inbetween the top four, reaching intermediate results. This behaviour is consistent with the tag-based experiment, reported in Section 6.4.1: the upper bounds of the ten sim_t measures in definition-based experiment have a correlation $\rho = .69$ ($p < .05$) with the tag-based similarity correlations.

When utilising GeReSiD as a gold standard, our lexical similarity measure can obtain high cognitive plausibility (upper bound $\rho \approx .74$), largely outperforming the basic WordNet-based measures applied directly to single terms (upper bound $\rho \approx .53$). In general, this technique tends to approximate semantic relatedness better than similarity, although the overall patterns are comparable. The top performance is reached with the following parameters: POS=*NN*, C=*Irish Indep*, sim_v=*com*, sim_t=$\{$*vector*, *path*, *lch*, *hso*$\}$). In such cases, the cognitive plausibility ρ falls in the interval $[.61, .74]$ for similarity, and $[.64, .74]$ for relatedness, showing a clear correlation with the human-generated dataset GeReSiD.

In order to further assess the reliability of these empirical results, the semantic similarity judgements can be treated as categorical. In particular, the rankings generated by the lexical similarity measure and the human subjects can be transformed into two categories (similar/not similar), based on average similarity ranking. The pairs ranked above the average ranking are therefore considered to be similar, and all the others are not similar. This analysis, applied to one of the best combinations of parameters (POS=*NN*, C=*Irish Indep*, sim_v=*com*, sim_t=$\{$*vector*$\}$), results in the contingency table in Table 6.7. Fisher's exact test indicates high statistical significance ($p < .001$), confirming

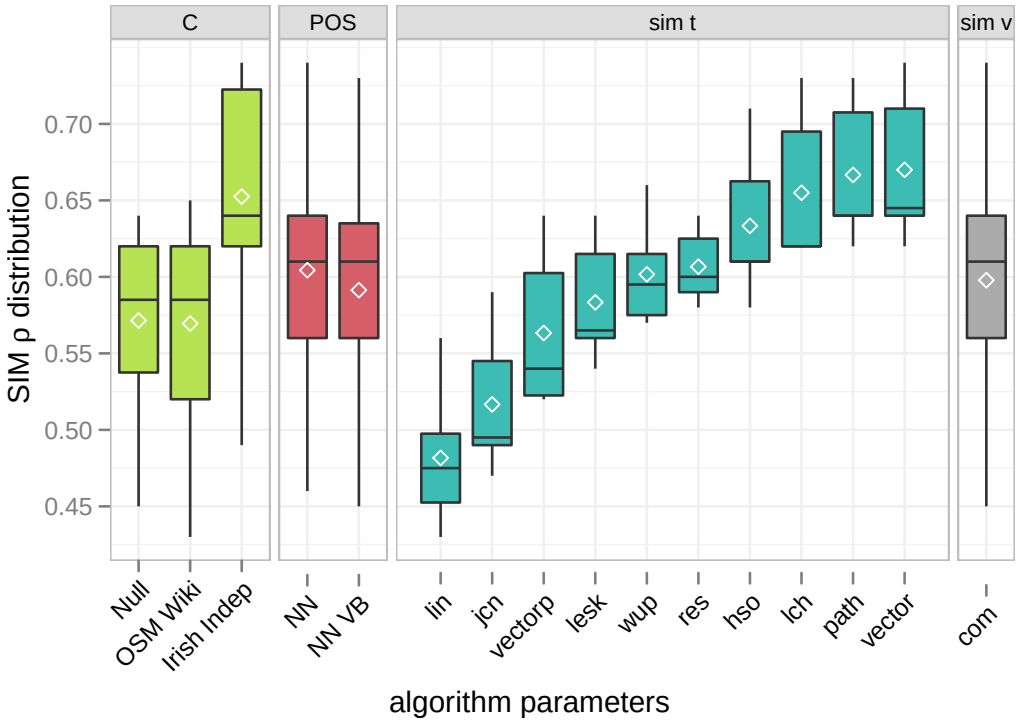

Figure 6.5: Lexical NetLexSiM: cognitive plausibility against GeReSiD for semantic similarity.

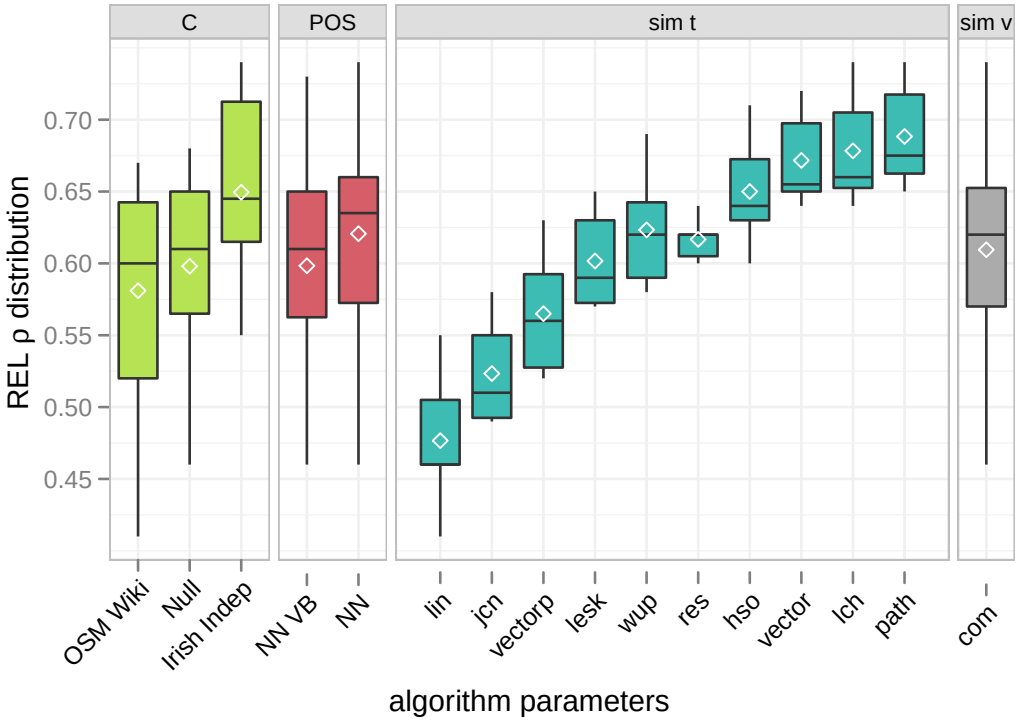

Figure 6.6: Lexical NetLexSiM: cognitive plausibility against GeReSiD for semantic relatedness.

		Lex Sim		Row
		Non-sim	Sim	total
Humans	Non-sim	21	4	25
	Sim	4	21	25
	Column total	25	25	50

Table 6.7: Lexical similarity, contingency table of pairs judged similar (sim) or non-similar (non-sim) by the human subjects and by the computational measure on the GeReSiD dataset. Fisher's exact test: $p < .001$

the high cognitive plausibility of this component of NetLexSiM. This result is highly consistent with the categorical analysis of the network component of NetLexSiM (see Section 6.2).

To assess the cognitive plausibility of this approach more in detail, it is useful to observe specific cases showing high discrepancy with the human-generated similarity judgements in the set of 50 concept pairs. Focusing on the best case (POS=*NN*, C=*Irish Indep*, sim_v=*com*, sim_t=$\{$*vector*$\}$, with $\rho = .74$), it is possible to observe that the pair <*sea, island*> is ranked 24th by human subjects, and 8th by the computational model. The definitions of these two concepts have large overlap, but they are highly related (8th in the relatedness ranking), and not similar. In this case, the model mistakes relatedness for similarity.

Furthermore, <*battlefield, monument*> is ranked 10th by the human subjects, and only 36th by the model. The concepts' definitions share only one term (*military*), and the other terms do not increase their similarity. Analogously, the similarity of <*industrial landuse, landfill*> is underestimated, as it is ranked 21st by humans, and 47th by the model. The reason for this wide mismatch lies in the fact that the definition of landfill is extremely short ("Where waste is collected, sorted or covered over"), and does not contain terms that would allow the model to capture some degree of similarity with the context of industrial production and waste processing. This issue might be mitigated by extending very short definitions by accessing more detailed definitions, such as DBpedia.

These results confirm our two initial hypotheses, i.e. (i) the lexical definitions of OSM concepts allows the computation of a more plausible semantic similarity than the OSM tags, and (ii) our knowledge-based BOW approach reaches high cognitive plausibility for both semantic similarity and relatedness. This empirical evidence, when compared with the pilot evaluation of Section 5.6, shows similarities and differences. The next section, via a statistical meta-analysis, compares these results with the preliminary experiment with the MDSM evaluation dataset, drawing conclusions on the effectiveness of the lexical component of NetLexSiM, and providing a set of guidelines to utilise our approach to semantic similarity.

Param Name	Param Value	*Similarity meta-analysis* MDSM dataset $\bar{\rho}$	GeReSiD SIM ρ	Meta SIM ρ	GeReSiD REL ρ
sim_v	com	.81	.74	.76 ± .03*	.74
	fes	.79	–	–	–
POS	NN	.81	.74	.76 ± .03*	.74
	NN VB	.81	.73	.75 ± .03	.73
	VB	–	–	–	–
C	Irish Indep	.79	.74	.75 ± .02*	.74
	OSM Wiki	.81	.65	.69 ± .07	.67
	Null	.81	.64	.68 ± .07	.68
sim_t	lch	.81	.73	.75 ± .03*	.74
	path	.77	.73	.74 ± .02	.74
	vector	.68	.74	.73 ± .03	.72
	hso	.76	.71	.72 ± .02	.71
	wup	.8	.66	.69 ± .06	.69
	res	.81	.64	.68 ± .07	.64
	vectorp	.66	.64	.64 ± .01	.63
	jcn	.77	.59	.63 ± .08	.58
	lin	.81	.56	.62 ± .11	.55
	lesk	.56	.64	.62 ± .03	.65
all	–	.77±.07	.68±.05	.7 ± .05	.68±.06

Table 6.8: Meta-analysis of lexical similarity experiments, including MDSM evaluation dataset and GeReSiD. (*) Best performance.

6.4.3 Lexical similarity meta-analysis

Our approach to lexical similarity has been evaluated on two datasets: the MDSM evaluation dataset (Section 5.6) and GeReSiD (Section 6.4.2). Using a meta-analytical approach, this section has the purpose of comparing these experiments, extracting a set of empirical guidelines to be adopted to obtain optimal results. A meta-analysis consists of a statistical analysis of results from individual studies to integrate the findings [93].

The cognitive plausibility obtained through these two evaluations shows common trends, but also a divergence for certain parameters. Restricting the discussion to semantic similarity, the overall performance of the approach obtains a median $\rho = .69$ for the MDSM evaluation dataset, and $.61$ for GeReSiD, with an upper bound respectively of $.81$ and $.74$. This fact is consistent with the evaluation of network similarity, in which co-citation approaches performed better on the MDSM dataset ($\rho \in [.55, .85]$) than on GeReSiD ($\rho \in [.5, .74]$) (see Section 6.3.3). This difference is due to the structure and coverage of the MDSM evaluation dataset (29 concepts structured in five sets) and GeReSiD (97 concepts in one set).

To obtain a representative average from these experiments, a meta-analytical approach is appropriate. Table 6.8 summarises the results of the meta-analysis, conducted with the Hunter-Schmidt method, based on a weighted mean [129]. The weights are the size of concept sample n, 29 for the MDSM evaluation dataset and 97 for GeReSiD. The meta-analysis is performed on semantic similarity, because we have carried out only one experiment on se-

mantic relatedness. The upper bounds for each parameter are combined into a weighted mean, with a 95% confidence interval. The upper bounds have strong correlation with the medians ($\rho = .97$ with $p < .001$), showing consistent results even when conducting the meta-analysis on the medians. In this sense, the upper bounds are not statistical outliers, and can be used as indicators of cognitive plausibility. For example, when corpus C is equal to *Null*, the similarity upper bound ρ for the MDSM dataset is $.81$, and $.64$ for GeReSiD, resulting in $.68 \pm .07$. The upper bound for semantic relatedness is $.68$.

While the overall trends in the two experiments on lexical similarity are consistent, the effect of individual parameters $\{POS, C, sim_t, sim_v\}$ varies. The POS filter determines what terms are included in the semantic vectors. Verbs in isolation (*VB*) do not show any correlation with human rankings, and were excluded. By contrast, nouns in isolation (*NN*) and combined with verbs (*NN VB*) obtain a close overall cognitive plausibility, respectively $.76 \pm .03$ and $.75 \pm .03$. In other words, the nouns convey the bulk of the semantic content, and the inclusion of verbs affects the results only marginally.

The corpora C are used to compute the semantic weights. The corpus extracted from the Irish Independent obtains the best overall plausibility ($.75 \pm .02$). On the other hand, the corpus based on the OSM Wiki website and the null corpus, i.e. uniform semantic weights, perform less well, obtaining comparable cognitive plausibility (respectively $.69 \pm .07$ and $.68 \pm .07$). This indicates that a corpus containing diverse natural language tends to outperform a corpus highly specialised on geographic concepts (OSM Wiki website), or uniform weights (*Null*). However, by looking at the confidence interval, all of the three corpora reach similar upper bounds, but the Irish Independent corpus shows less variability, and should therefore be preferred.

A crucial parameter is the vector-to-vector sim_v similarity technique. While *com* reaches the best results in both experiments, resulting in an overall plausibility of $.76 \pm .03$, *fes* performs less well with the MDSM dataset, and does not obtain statistically significant results with GeReSiD. For this reason, *com* by Corley and Mihalcea [54], is certainly the optimal option.

Finally, the ten term-to-term similarity functions sim_t have a major impact on the cognitive plausibility of the algorithm. All sim_t show positive correlation with the human dataset, but some measures perform considerably better than others. Moreover, a high variability can be noticed between the two experiments. For example, the *vector* measure obtains comparatively low cognitive plausibility against the MDSM evaluation dataset (upper bound $.68$), but outperforms all other measures in GeReSiD ($.74$).

Such high variability is not uncommon in the literature on semantic similarity. In a study by Budanitsky and Hirst [39], measures *jcn*, *hso*, *lin*, *lch*, and *lesk* obtain very different cognitive plausibility against two well-known similarity datasets (Rubenstein and Goodenough [252] and Miller and Charles [206]). For example, out of these five measures, *jcn* ranks first when compared with the former, and only fourth with the latter. When seemingly divergent results are obtained, the meta-analytical approach offers statistically sound techniques to combine them into a representative average. The meta-analysis

depicted in Table 6.8 indicates the cognitive plausibility of the ten WordNet-based sim_t for semantic similarity of OSM concepts.

The measures that reach the top overall performance are *lch*, *path*, *vector*, and *hso*, with upper bounds in range $[.72, .75]$. Two measures, *wup* and *res*, obtain intermediate cognitive plausibility, in the interval $[.68, .69]$. The other four measures, *vectorp*, *jcn*, *lin*, and *lesk* rank lower, falling in the interval $[.62, .64]$. Analysing the specific characteristics of the measures, it is possible to point out general trends. The measures exploiting information content (*res*, *jcn*, *lin*) obtain generally low results. The three measures based on the glosses can obtain comparatively high plausibility (*vector*), or low (*vectorp* and *lesk*). By contrast, the measures that rely on the shortest paths, without information content and concept glosses obtain the most plausible results (*path*, *lch*, *hso*, *wup*). This indicates that, although more complex measures can obtain optimal results in certain contexts, simpler shortest path-based measures tend to perform more reliably across the two datasets.

The ultimate purpose of this body of empirical evaluations is to provide practical advice for obtaining optimal results when computing semantic similarity or relatedness of OSM geographic concepts with NetLexSiM. Based on the meta-analysis reported in Table 6.8, we can indicate a set of guidelines in the selection of parameters in the process of comparing concepts on a lexical basis:

- **POS filter**: in lexical definitions, nouns (*NN*) convey most semantic similarity. The addition of verbs slightly affects the results, at times for the better, at times for the worse.
- **Corpus** C: to obtain the highest cognitive plausibility, a general text corpus (extracted from the *Irish Independent* newspaper) is preferable both to a domain specific dataset (OSM Wiki website), and to uniform weights (*Null* corpus).
- **Term-to-term measure** sim_t: Out of the ten being analysed, three measures appear to be the best choice to compute semantic similarity between geographic classes: *lch* (Leacock and Chodorow [170]), *path* (Rada et al. [239]), and *vector* (Patwardhan and Pedersen [232]). Although lexical-chain measure *hso* by Hirst and St-Onge [124] also obtains promising results, a high temporal complexity makes its usage problematic (see Section 5.6.2).
- **Vector-to-vector measure** sim_v: The vector-to-vector similarity *com* by Corley and Mihalcea [54] largely outperforms *fes* (Fernando and Stevenson [75]), while the opposite is true in the context of paraphrase detection.

To conclude, the following combination of parameters is the best policy that we propose to compute lexical semantic similarity for geographic concepts:

POS=*NN*,	C=*Irish Indep*,	sim_t={*lch*, *path*, *vector*}	sim_v=*com*

These particular combinations consistently show the best overall cognitive plausibility, and therefore can be recommended as a general policy to compute lexical semantic similarity between geographic classes. While the network-based measures clearly approximate semantic similarity better than relatedness (see Section 6.3.3), the trend is less clear-cut in the case of the lexical measures, which obtain similar results for both similarity and relatedness. Further research is needed to identify lexical techniques specialised in similarity and relatedness of volunteered geographic concepts in OSM. The next section moves on to evaluate the combination of the two similarity components, network and lexical, into a hybrid similarity for OSM concepts, reaching higher cognitive plausibility.

6.5 NetLexSiM: evaluation of hybrid similarity

To compute plausible semantic similarity and relatedness for OSM concepts, NetLexSiM takes two aspects into account: the topological similarity in the OSM Semantic Network (network similarity s_{net}), and the similarity of concept definitions (lexical similarity s_{lex}). The structure of NetLexSiM was outlined in Section 3.6. In Section 3.8 two methods to combine these aspects into a hybrid measure were presented: a score combination s_{sc}, and a rank combination s_{rk}.

This section describes an empirical evaluation of these two techniques, showing that the cognitive plausibility of such hybrid measures is generally higher than the individual network and lexical measures. An experiment is set up in Section 6.5.1, and its results are discussed in Section 6.5.2.

6.5.1 Experiment setup

To explore the effectiveness of score and rank combination methods, an experiment was set up using GeReSiD as a gold standard. The most plausible cases were selected for network and lexical relatedness and similarity measures, based on the meta-analyses described in Sections 6.3.3 and 6.4.3. As we are interested in assessing whether the combination methods are able to improve the results at the top of the range, the selection is restricted to the top 30 cases for both approaches, which we consider to be a representative sample of the network and lexical measures. As discussed above, the top cases are not statistical outliers, but accurately reflect general trends in the cognitive plausibility. The experiment was set up with the following input parameters:

- Combination methods: score combination s_{sc}, and rank combination s_{rk}.
- Combination factor α: ten discrete equidistant levels $\in [0, 1]$. When $\alpha = 0$, only the lexical measure is considered; $\alpha = 1$, on the other hand, corresponds to the network measure.
- Network similarity s_{net}: 30 most cognitively plausible cases when compared with GeReSiD.

			Score comb s_{sc}		Rank comb s_{rk}	
		α	max ρ	success %	max ρ	success %
SIM	*lex*	0	.74	–	.74	–
	hyb	.2	.79	74.4	.79	73.1
	hyb	.4	.81	87.5	.83	100.0
	hyb	.5	.82*	91.9	.84*	100.0
	hyb	.6	.81	95.6	.83	100.0
	hyb	.8	.8	96.9	.79	86.9
	net	1	.73	–	.73	–
REL	*lex*	0	.74	–	.74	–
	hyb	.2	.78*	91.9	.78	95.0
	hyb	.4	.78*	96.2	.81*	95.6
	hyb	.5	.78*	98.8	.8	91.2
	hyb	.6	.77	96.9	.78	89.4
	hyb	.8	.75	71.2	.72	45.0
	net	1	.64	–	.64	–

Table 6.9: Cognitive plausibility of hybrid measures. Max ρ is the upper bound obtained by an approach. *net*: network measure; *lex*: lexical measure; *hyb*: hybrid measure. (*) Best performance. For all Spearman's tests, $p < .001$

– Lexical similarity s_{lex}: 30 most cognitively plausible cases when compared with GeReSiD.

For each value of α, each case of s_{net} and s_{lex} were combined through s_{sc}, and s_{rk}. This resulted in the cognitive plausibility of two sets of hybrid measures, 18,000 for similarity, and 18,000 for relatedness (with $p < .001$ for all Spearman's correlation tests). A hybrid measure is considered *successful* if it outperforms both its components s_{net} and s_{lex}, i.e. the cognitive plausibility of the hybrid measure is higher than network and lexical similarity, formally $\rho_{hyb} > \rho_{net} \wedge \rho_{hyb} > \rho_{lex}$. If the hybrid measure is lower than any of its components, it has failed.

6.5.2 Experiment results

Clear patterns emerge from the experiment results. Hybrid measures, combining network and lexical similarity, show a consistent advantage over individual measures. Moreover, the ranking combination s_{rk} performs consistently better than the score combination s_{sc}, obtaining higher plausibility and success rate. The results for semantic similarity are generally higher than those for relatedness, a trend that has been confirmed throughout this chapter.

Table 6.9 summarises the experiment results, contrasting the upper bound (maximum ρ) obtained by *net* and *lex* measures in isolation, and *hyb* when combined. The cognitive plausibility of hybrid measures is substantially higher than the individual measures, with a peak at $\rho = .84$ for similarity ($\alpha = .5$), and $\rho = .81$ for relatedness ($\alpha = .4$). This empirical evidence points out that the optimal value of α tends to fall in the interval $[.4, .6]$, drawing information evenly from the network and lexical components. The cognitive

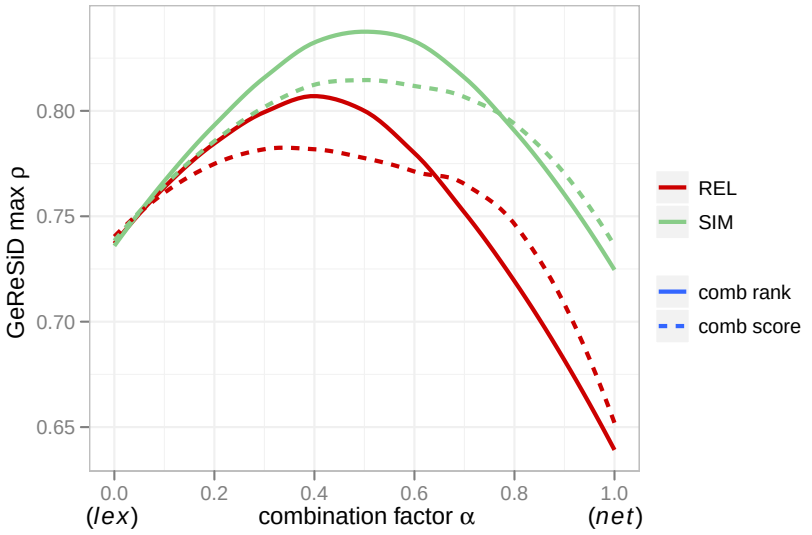

Figure 6.7: Cognitive plausibility of hybrid measures, including network and lexical relatedness (REL) and similarity (SIM). The lines are smoothed using a linear model; *comb rank*: rank combination s_{rk}; *comb score*: score combination s_{sc}; $\alpha \in [0, 1]$. Each correlation test includes 50 pairs of rankings without ties.

		Hyb Sim		Row
		Non-sim	Sim	total
Humans	Non-sim	[19,23]	[2,6]	25
	Sim	[2,6]	[19,23]	25
	Column total	25	25	50

Table 6.10: Hybrid similarity measure, summary of 1,800 contingency tables of pairs judged similar (sim) or non-similar (non-sim) by the human subjects and by NetLexSiM on the GeReSiD dataset, with $\alpha = .5$ and s_{rk}. The table contains the ranges of each of the four categorical combinations. Fisher's exact test: $p < .001$

plausibility of similarity is skewed towards network measures, while relatedness is best approximated when favouring the lexical side.

In order to provide further evidence of the reliability of these results, we have conducted an analysis treating the semantic similarity judgements as categorical. Considering the 1,800 cases with one of the optimal values of α (.5) and merging the network and lexical similarity with ranking combination (s_{rk}), the judgements generated by the human subjects and by the hybrid measures were transformed into two categories (similar/not similar). All the pairs ranked above the average rank are considered to be similar, and the others are not similar. Applying this analysis to these 1,800 cases, we obtained highly consistent contingency tables, summarised in Table 6.10. Using Fisher's exact test, all of these 2x2 contingency tables obtain high statistical significance ($p < .001$). This result confirms the high cognitive plausibility of NetLexSiM, confirming the results outlined in Sections 6.3.2 and 6.4.2.

The success rate, expressed as a percentage, indicates how many times a hybrid measure outperformed both individual measures. As it is possible

to notice in Table 6.9, when $\alpha \in [.4, .6]$, the success rate is very high, in the interval $[87.5\%, 100\%]$. In particular, the rank combination s_{rk} outperforms *all* individual measures (100%). High success rates are also observable when $\alpha \in (0, .4)$, with an average success rate of 82.9%. At the other end of the spectrum ($\alpha \in (.6, 1)$), the average success rate is 75%. In none of the cases under consideration, a hybrid measure was lower than both components. The success rates reported in Table 6.9 show that, overall, hybrid measures are a preferable choice to compute semantic similarity and relatedness for geographic concepts. In particular, when using ranking combination s_{rk} with optimal values of α, the hybrid measures obtain a success $> 89\%$, both for semantic similarity and relatedness.

The performance of hybrid measures is depicted in Figure 6.7, highlighting the impact of α on the combined similarity for similarity and relatedness, adopting the two combination techniques (s_{sc} and s_{rk}). The roughly bell-shaped curves in the Figure display the benefit of the hybrid measures ($\alpha \in (0, 1)$) over the individual measures, at the extremes of the horizontal axis ($\alpha = 0$ corresponds to lexical measures, $\alpha = 1$ to network measures). While for semantic similarity the plot is symmetrical, semantic relatedness is skewed towards lexical similarity. This is due to the fact that, for relatedness, lexical measures obtain higher cognitive plausibility than network measures. By contrast, for similarity, the initial performance of the components is comparable ($\approx .74$), giving a certain symmetry to the behaviour of hybrid measures. From the empirical evidence presented in this section, the following conclusions can be drawn:

- The hybrid measures, combining network-based and lexical semantic similarity, tend to outperform individual measures, obtaining higher cognitive plausibility. Compared with the upper bounds for network measures ($\approx .73$ for similarity, $.64$ for relatedness) and lexical measures ($.74$ both for similarity and relatedness), hybrid measures reach an upper bound of $\rho = .84$ for similarity, and $\rho = .81$ for relatedness.
- Of the two combination methods introduced in Section 3.8, ranking combination s_{rk} obtains more promising results than direct score combination s_{sc}. This difference notwithstanding, both combination techniques obtain better plausibility than individual measures.
- Both components, network-based and lexical, contribute to increasing the cognitive plausibility of the hybrid measures. This is indicated by the fact that the optimal value of α falls in the interval $[.4, .6]$.
- With optimal values of α and ranking combination s_{rk}, hybrid measures outperform individual measures in the vast majority of the cases under consideration, obtaining a consistently high success rate ($> 89\%$).

The hybrid approaches evaluated in this section seem the most suitable way to compute plausible similarity and relatedness measures for OSM concepts. Such hybrid measures show the complementary nature of network and

lexical measures, overcoming the limitations of each individual approach. Indeed, they should be preferred to individual network and lexical measures. General conclusions from this body of empirical evidence will be discussed in Chapter 7.

Our measure for concept similarity, NetLexSiM, operates at the concept level. An alternative approach to geo-semantic similarity that we explored is that of the holistic similarity of viewports, as opposed to individual concepts or map features (see Section 3.9). Having evaluated NetLexSiM, we move on to outline an evaluation of this holistic approach to viewport similarity, which treats map viewports as documents in a information retrieval (IR) system.

6.6 Evaluation of viewport similarity

While most semantic similarity approaches focus on individual classes or instances of geographic entities, in Section 3.9 we have proposed a viewport-based, holistic Geographic Information Retrieval (GIR) system, published in [23]. This system aims at capturing the overall semantics of a viewport, treating viewports in a manner similar to documents in text-based IR. To capture a holistic impression of semantics in a web map, we have defined semantic descriptors as vectors D_v that represent the overall semantic content of the viewport v in which the map is displayed. This section depicts a case study with our viewport information retrieval system, suggesting possible applications, strengths and weaknesses of the approach.

This approach to GIR draws from the observation that certain tasks that users perform on Web maps are not only analytical, i.e. focused on the decomposition of large objects into simpler parts, but are also holistic, treating a geographic area as a unified entity. Analytical tasks involve, for example, the examination of specific target objects. On the other hand, examples of spatial holistic tasks are the classification of an urban area versus a rural area, in which the user needs not to focus on individual objects, but classify the area as a whole. These holistic tasks should not be considered in opposition to analytical tasks, but they are intertwined in the complex cognitive interplay that occurs in the interaction with information retrieval systems. To provide an initial evaluation of the approach, we observe its behaviour on a sample of viewports, extracted from the Irish OSM dataset.

6.6.1 Sample viewports

In this case study we consider a small set of viewports selected from a bounding box bb, corresponding to the surroundings of Dublin. This geographic area g contains a total of $\approx 20,000$ features. Five sample viewports were extracted at zoom level $z = 15$, including a Dublin suburb, a park, a port, and two seaside towns. Five semantic descriptors D_v were then computed for each of the six viewports v, linear, logarithmic, information-theoretic, surface, and a mean of the first four, limiting for the sake of illustration the number of feature types

to ten out of the 64 visible types. The self-information of each feature type was computed on g, and not on the entire OSM dataset. The viewports and corresponding descriptors are reported in Table 6.11.

Looking at the weights of the descriptors, it is possible to trace behaviour, and pros and cons of each of the four weighting mechanisms proposed in Section 3.9. The linear weights reflect the number of features visible in the viewport. For this reason, important features such as the Phoenix Park in v_2 rank very low, and roads rank very high in most viewports. This is because roads are represented in numerous small chunks, while large objects, such as parks, consist of one large polygon, and the linear weights do not take this aspect into account.

This problem is partly addressed by the logarithmic weights, which smooth the results by increasing the importance of types with few features and by decreasing that of types with many occurrences in the viewport. In v_3, the logarithmic weight of 'building' has been doubled, while that of 'coastline,' another type of feature modelled in small chunks, has been reduced. However, despite the smoothing, the resulting weights are still strongly biased towards types such as 'road,' and 'building,' while large and infrequent features are squeezed into small weights.

The area-based technique tends to correct this bias. In v_2, the type 'park,' which is almost ignored by the previous weighting mechanisms, is the most important in the viewport. Thanks to its large area, this type gains a lot of influence in the descriptor. As it is possible to notice by the frequent 0 values in the area column, only a subset of features are polygons and can be included in this descriptor, resulting in a limited information problem.

The behaviour of the self-information weights is more difficult to interpret. For types that occur very frequently in g, such as 'road' and 'building,' $I(t)$ is low (3.51 and 4.85), while less frequent types have higher $I(t)$ (12.21 for 'port,' and 11.56 for 'park'). This is correct, but when the self-information values are multiplied by the number of features, they smooth the results to a limited extent. In the case of viewport v_2, according to the self information weights, buildings are more important than parks, maintaining the bias of the linear and logarithmic weights. As is expected, the mean weights are heavily smoothed, but maintain some of the bias of the linear, logarithmic, and self-information weights.

In holistic geographic information retrieval, the user can retrieve, instead of specific geographic features, viewports that visually represent geographic areas rendered at a given zoom level. In this case, the user's information need is not a specific spatial information, e.g. where is the target object, or what is the area of the target object, etc., but is an implicit semantic judgement on viewports displayed on the screen. A viewport representing a geographic area that fulfills the user's information need, e.g. a seaside town or a commercial port, is used as a query viewport to retrieve viewports conveying similar semantic content. To achieve this, the system has to be able to model this *implicit judgement* on geographic content displayed in a viewport v.

Viewport	Feat Type	*Weighting schemes* Linear	Log	Area	Self-Info	Mean
v_1	building	.495	.321	.162	.388	.341
	coastline	–	–	–	–	–
	commercial	.029	.113	.034	.058	.058
Milltown: suburb of Dublin, with hospital, college, and residential estates.	hospital	.01	.056	.111	.021	.05
	industrial	–	–	–	–	–
	park	.01	.056	.043	.025	.033
	port	–	–	–	–	–
	road	.427	.309	–	.462	.3
	town	.01	.056	.59	.021	.169
	wood	.019	.089	.06	.025	.048
v_2	building	.229	.272	.035	.156	.173
	coastline	–	–	–	–	–
	commercial	–	–	–	–	–
Phoenix Park: Large urban park with zoo, polo grounds, and the American embassy.	hospital	–	–	–	–	–
	industrial	–	–	–	–	–
	park	.029	.086	.789	.064	.242
	port	–	–	–	–	–
	road	.257	.285	–	.242	.196
	town	–	–	–	–	–
	wood	.486	.358	.175	.539	.389
v_3	building	.087	.16	.089	.047	.096
	coastline	.442	.268	–	.542	.313
	commercial	–	–	–	–	–
Howth: seaside town with tourist attractions, cliffs, and trekking trails.	hospital	–	–	–	–	–
	industrial	–	–	–	–	–
	park	.029	.096	.276	.052	.113
	port	–	–	–	–	–
	road	.308	.243	–	.233	.196
	town	.01	.048	.309	.015	.095
	wood	.125	.184	.325	.111	.186
v_4	building	.296	.207	.158	.175	.209
	coastline	.194	.183	–	.257	.158
	commercial	.143	.165	.175	.214	.174
Dun Laoghaire: seaside town with a small port, a private school, and a hospital.	hospital	.01	.042	.088	.017	.039
	industrial	.02	.067	.026	.028	.035
	park	.01	.042	.018	.02	.022
	port	.01	.042	.184	.021	.064
	road	.306	.209	–	.251	.191
	town	.01	.042	.351	.017	.105
	wood	–	–	–	–	–
v_5	building	.17	.19	.12	.08	.14
	coastline	.136	.176	–	.144	.114
	commercial	.386	.244	.326	.461	.354
Dublin Port: docks of the Dublin port, where large ships load and unload containers.	hospital	–	–	–	–	–
	industrial	.227	.209	.174	.251	.215
	park	–	–	–	–	–
	port	.011	.048	.38	.019	.115
	road	.068	.133	–	.044	.062
	town	–	–	–	–	–
	wood	–	–	–	–	–

Table 6.11: Viewport semantic descriptors D_v for 5 sample viewports, with weights computed using four approaches, and their mean. Symbol '–' corresponds to 0.

Viewport Name	*	Cosine					Euclidean				
		v_1	v_2	v_3	v_4	v_5	v_1	v_2	v_3	v_4	v_5
Milltown	v_1	–	–	–	–	–	–	–	–	–	–
Phoenix Park	v_2	.55	–	–	–	–	.52	–	–	–	–
Howth	v_3	.54	.65	–	–	–	.55	.59	–	–	–
Dun Laoghaire	v_4	.82	.38	.68	–	–	.72	.48	.66	–	–
Dublin Port	v_5	.37	.15	.29	.73	–	.46	.35	.45	.68	–

Table 6.12: Similarity of sample viewports. The matrices are symmetrical, and their diagonal values are equal to 1. The euclidean similarity has been computed as $1 - d$, where d is the euclidean distance.

6.6.2 Sample viewport similarity

The semantic similarity of the viewports can be computed via the cosine distance between their descriptors, as shown in Table 6.12. In this case, the two vector similarity measures rank the pairs in the same way. The pairs with the highest similarities are $\langle v_1, v_4 \rangle$, and $\langle v_4, v_5 \rangle$. Viewports v_1 and v_4 are semantically very similar, and this result is satisfactory.

The pair $\langle v_4, v_5 \rangle$ is surprising, because a tourist seaside town appears semantically very different to a commercial port. This result is easily explained by the absence of feature types that would increase the distance between the two viewports, such as restaurants, tourist attractions, amenities, which are strongly present in v_4 but not in v_5. The two viewports share the presence of the seaside, a port, many buildings and commercial activities, all aspects that are captured by the ten feature types considered in this case study.

On the other hand, the least similar pairs are $\langle v_2, v_5 \rangle$, and $\langle v_3, v_5 \rangle$. This seems to be a valid result, as these pairs represent very different areas, sharing very few feature types. Based on this case study, it can be concluded that the holistic semantic descriptors proposed in this paper are a promising approach to semantic similarity. However, an extensive evaluation is necessary to assess the cognitive plausibility of this approach to computing the semantic similarity of viewports.

Adopting this approach to IR, the user can submit a relevant viewport to the system, and retrieve semantically similar viewports. Such a holistic exploration of geo-data has useful applications in the context of tourism, urban planning, and property development, in which users are often interested in retrieving and comparing similar geographic areas. Plans for further development of this holistic viewport GIR system are outlined in Chapter 7.

6.7 Summary

This chapter covered the evaluation of the Network-Lexical Similarity Measure (NetLexSiM), our measure to compute the semantic similarity of OpenStreetMap (OSM) concepts, tapping the OSM Semantic Network. First, having reached promising results in the pilot evaluation in Chapter 5, we proceeded

to develop and validate a novel gold standard, the Geo Relatedness and Similarity Dataset (GeReSiD) (Section 6.2). This human-generated dataset focuses on OSM concepts, and distinguishes between semantic similarity and relatedness. GeReSiD was then utilised as a gold standard to further evaluate the co-citation network similarity measures, obtaining results highly consistent with the pilot evaluation, with an upper bound $\rho = .74$ for similarity, and $.64$ for relatedness (Section 6.3).

The lexical measure was also evaluated against GeReSiD, obtaining a performance of $\rho \approx .74$ for both similarity and relatedness (Section 6.4). The two components were combined into a hybrid measure, which obtain the best performance, $\rho = .84$ for similarity, and $.81$ for relatedness (Section 6.5). This set of empirical evidence was used to indicate general guidelines to compute semantic similarity and relatedness in the context of OSM, which can provide semantic support for data mining, information integration, and map personalisation.

Finally, the chapter focused on the evaluation of our approach to viewport semantic similarity, treating map viewports as documents in a information retrieval (IR) system (Section 6.6). The next chapter will draw conclusions about the body of work we have presented in this thesis. The merits and limitations of the OSM Semantic Network, NetLexSiM and the other contributions are summarised, indicating promising directions for future work.

CHAPTER

SEVEN

CONCLUSIONS

7.1 Thesis summary

The main endeavour of this thesis is to provide semantic support for the leading spatial crowdsourcing project OpenStreetMap (OSM), bridging the wide gap between Volunteered Geographic Information (VGI) and the Semantic Geospatial Web. In particular, we started by merging knowledge from open knowledge bases with OSM, using heuristics and matching techniques. The issue encountered in this process highlighted a semantic gap in the OSM semantic model, i.e. the lack of a rich, machine-readable knowledge representation for concepts. In OSM, the meaning and nature of the geographic concepts (called 'tags' in the project jargon) are negotiated on the OSM Wiki website. The project's shared semantic ground emerges from this crowdsourced process, in which contributors propose, amend, and describe concepts.

In order to tap this rich and evolving knowledge repository, we developed the OSM Semantic Network, a machine-readable structure extracted from the OSM Wiki website.[1] As part of this semantic support for OSM, we released the OSM Wiki Crawler, an open source tool. The crawler scans the OSM Wiki website and encodes the semantic knowledge defined by the project contributors into the OSM Semantic Network.[2] Subsequently, we focused on the computation of the semantic similarity of OSM concepts. Our Network-Lexical Similarity Measure (NetLexSiM) combines network and lexical similarity measures, tapping the knowledge encoded in the OSM Semantic Network. This measure can provide semantic support for a number of advanced tasks in Geographic Information Retrieval (GIR), map personalisation, and information integration. To evaluate NetLexSiM, we designed and collected the Geo Relatedness and Similarity Dataset (GeReSiD), which can be used as a human-generated gold standard in geo-semantic research.

This thesis is structured in seven chapters. In Chapter 1, we introduced the core research problem that we addressed, i.e. the semantic gap in OSM, in the emerging framework of VGI, between the two worlds of Web 2.0 and the Semantic Web (Section 1.2). The aims and hypotheses underlying our efforts were defined, highlighting the need for semantic similarity measures for OSM (Section 1.3). The chapter outlined the thesis contribution and its structure

[1] http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork
[2] http://github.com/ucd-spatial/OsmWikiCrawler

(Sections 1.4 and 1.5).

Chapter 2 surveyed in detail the vast scientific literature relevant to our work. Our preliminary work, in which the OSM semantic gap was identified, addressed user profiling and personalisation (Section 2.2). The rise of spatial crowdsourcing and VGI was subsequently discussed (Section 2.3). The area of geo-semantics draws from a broad scientific literature, ranging from linguistics, computer and cognitive science (Section 2.4). Interdisciplinary research on semantic similarity was then surveyed, identifying issues of particular relevance to the context of OSM geographic concepts (Section 2.5).

Chapter 3 outlined the approaches we developed as a contribution to the semantic model of OSM. As discussed above, the semantic research problem was identified in our preliminary work, conducted in collaboration with the National University of Ireland, Maynooth, in the framework of the Strategic Research in Advanced Geotechnologies (StratAG) cluster (Section 3.2). This preliminary work included the RecoMap recommender system, which highlights geographic features based on users' behaviour. In order to semantically enrich the OSM dataset, we devised a technique to enrich the OSM dataset with DBpedia and LinkedGeoData.

As a core semantic support tool to OSM, we developed the OSM Semantic Network, an open source semantic resource (Section 3.3), supporting the semantic model of OSM (Section 3.4). The OSM Semantic Network be used to compute the semantic similarity and relatedness of geographic concepts (Section 3.5). NetLexSiM, our approach to computing the semantic similarity of OSM concepts, rests on two complementary pillars: network-based, co-citation topological measures (Section 3.6) and a knowledge-based measure to compare the lexical definitions of concepts (Section 3.7), combined in a hybrid similarity measure (Section 3.8). An alternative approach to semantic similarity, which considers map viewports as semantic units, was explored in our GIR system (Section 3.9).

Chapter 4 details aspects of the implementation of our contributions, engaging with current, state-of-the-art web technologies. In order to extract the OSM Semantic Network, we developed an open source tool, the OSM Wiki Crawler (Section 4.2). In the context of our preliminary work on map personalisation, we developed a Web platform for map personalisation and visualisation. In the design process, we conducted a user survey, asking open source contributors and developers their opinion about a set of Web mapping tools, exploring the evolving landscape of free and open-source software (FOSS) (Section 4.3). The resulting Web platform for map personalisation and visualisation hosts Web services tailored to exploit implicit feedback mechanisms to provide a range of personalisation services (Section 4.4). This platform helped identify the core semantic gap in OSM data, which informed the core contributions of this thesis.

Chapter 5 reports on the preliminary empirical evaluation of our contributions, starting from the evaluation of our approach to enriching the semantics of OSM with DBpedia (Section 5.2). To evaluate the effectiveness of NetLexSiM, we studied its cognitive plausibility (Section 5.3). A pilot evalu-

ation of NetLexSiM was conducted on an existing gold standard, the MDSM evaluation dataset (Section 5.4). On this dataset, we performed a pilot evaluation for the network-based component of NetLexSiM (Section 5.5), followed by the lexical component (Section 5.6). The analogy of the 'similarity jury' was then discussed and evaluated (Section 5.7).

Chapter 6 focuses on an evaluation of NetLexSiM, tailored specifically for the OSM semantic model. We developed the Geo Relatedness and Similarity Dataset (GeReSiD) (Section 6.2), and used it to evaluate NetLexSiM for network (Section 6.3), lexical (Section 6.4) and hybrid similarity (Section 6.5), obtaining the highest cognitive plausibility with the hybrid approach. Meta-analyses are utilised to draw general conclusions from the comparison between these experiments and the pilot evaluation of Chapter 5. A preliminary evaluation of our approach to holistic viewport similarity concludes the chapter (Section 6.6).

The remainder of this chapter is organised as follows. First, we state the original scientific contribution presented in this thesis (Section 7.2). We then address in detail the limitations of our work, pointing out promising solutions (Section 7.3). Finally, we illustrate directions for future work in the area of semantics for OSM and VGI (Section 7.4).

7.2 Thesis contribution

The original scientific contributions described in this thesis aim at providing semantic support for the spatial grassroots project OpenStreetMap (OSM). Such research efforts were conducted in two phases: (1) a preliminary investigation of implicit feedback for map personalisation, and (2) the development of semantic support tools OSM, including the OSM Semantic Network, and NetLexSiM. The contributions developed throughout these two phases are discussed in Sections 7.2.1 and 7.2.2. For each contribution, we summarise the key findings and salient conclusions.

7.2.1 Preliminary work

First, we explored the area of map personalisation, and implicit feedback. This work was conducted in collaboration with other research projects, in the framework of the cluster Strategic Research in Advanced Geotechnologies (StratAG). The contributions originated from this preliminary phase are the following:

1. *RecoMap*: a spatial recommender system based on implicit feedback analysis, which monitors user interaction and highlights relevant features (Section 3.2.1).

2. A Web platform for map personalisation and visualisation: an architecture to implement personalisation and implicit feedback services (Section 4.4).

3. An online survey on free and open-source software (FOSS) projects, including Web GUI toolkits, spatial DBMSs, and Web development frameworks, to investigate their strengths and weaknesses from the users' perspective (Section 4.3). The survey gives the following indications:
   - The responders expressed high satisfaction with support for existing standards, stability, interoperability with other systems, and community support.
   - The responders reported intermediate satisfaction in relation to performance, scalability, and extendibility of the projects, and lower satisfaction for user friendliness, quality of the documentation, and the frequency of updates.
   - Projects Grails, PostGIS, MooTools, MapServer, and OpenLayers obtained the best overall user scores.

To model user profiles for personalisation and implicit feedback, we identified an open problem. While objects can be considered as isolated semantic units, their relationships with other objects can increase the expressiveness and inferential power of a profiling techniques. In other words, beyond the specific feature types, there are implicit semantic connections, whose knowledge can support a number of advanced spatial tasks. For example, if a user repeatedly queries a retrieval system to see schools and universities, her profile should model an interest in education. This piece of information can then be used to perform more accurate recommendations, search results, and group profiling. The exploration of this issue has resulted in a range of contributions to the area of geo-semantics, with a particular focus on the OSM semantic model, summarised in the next section.

7.2.2 Contributions to geo-semantics

The core contributions of this thesis are aimed at providing semantic support for OSM, the leading VGI project. First, we devised a technique to enrich OSM with knowledge from DBpedia. We then developed a novel semantic network, and a measure for semantic similarity for geographic concepts, NetLexSiM, exploring network and lexical approaches to semantic similarity. In summary, the core contribution to knowledge in our thesis consists of the following aspects:

1. OSM semantic enrichment with open knowledge bases: a technique to enrich OSM semantics that retrieves corresponding geographic entities in open geo-knowledge bases, such as OSM and DBpedia (Section 3.2).
2. The OSM Semantic Network: a crowdsourced semantic network representing OSM geographic concepts, extracted in a bottom-up process from the OSM Wiki website, released as Open Knowledge (Section 3.3).[3] The

[3] `http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork`

dedicated crawler we have developed for this purpose is released under a GPL license, and can be used to extract up-to-date versions of the network over time.[4]

   - The OSM Wiki website contains an informal conceptualisation of geographic concepts, used by OSM contributors to describe the geographic reality that the map is supposed to represent.
   - The OSM Semantic Network captures this conceptualisation in a machine-readable structure, which can support the usage of OSM vector data.

3. Geo Relatedness and Similarity Dataset (GeReSiD): as part of our contribution to geo-semantic research, we collected a human-generated dataset for semantic similarity and relatedness. This dataset allows the assessment of the cognitive plausibility of measures when applied to OSM concepts, determining empirically to what degree a measure approximates relatedness and similarity (Section 6.2). These are the findings obtained from the survey:

   - Human subjects strongly agree on cases of very high and low semantic relationships, and tend to have lower agreement on the intermediate cases.
   - Semantic relatedness and similarity are strongly correlated ($\rho = .95$). Semantic relatedness scores consistently higher than similarity, confirming the more specific nature of semantic similarity.

4. Network-Lexical Similarity Measure (NetLexSiM): a novel approach to computing the semantic similarity of OSM concepts, i.e. 'tags' in the OSM terminology. NetLexSiM exploits two aspects of the OSM Semantic Network to compare geographic concepts semantically: network-based similarity and lexical similarity techniques (see next points for more details). Relying on a body of empirical evidence, we provide a set of guidelines to compute the semantic relatedness and similarity for OSM, using NetLexSiM. Sets of pre-computed similarity scores for the entire OSM Semantic Network are available online.[5]

5. Network NetLexSiM: the network-based component of NetLexSiM relies on existing network co-citation algorithms such as SimRank and P-Rank (Section 3.6). The following conclusions can be drawn from the empirical evaluation reported in Sections 5.5 and 6.3:

   - Recursive co-citation algorithms reach high cognitive plausibility: SimRank [136] obtains $\rho = .7 \pm .08$ for similarity and .55 for relatedness, while P-Rank [318] reaches $.74 \pm .01$ for similarity, .65 for relatedness. Among the non-recursive algorithms, classic Co-citation [271] reaches the highest plausibility (.64 for similarity, .49 for relatedness).

[4] http://github.com/ucd-spatial/OsmWikiCrawler
[5] http://spatial.ucd.ie/osn/similarities

- The best results are obtained when similarity flows across the network edges with slow decay ($C \in [.8, .9]$), and favour incoming links, as opposed to outgoing links ($\lambda \in [.9, 1]$).

6. Lexical NetLexSiM: the lexical component of NetLexSiM consists of a knowledge-base similarity measure, which combines existing WordNet and paraphrase detection techniques. The measure is applied to lexical definitions of concepts (Section 3.7). The empirical evaluation in Sections 5.6 and 6.4 gives guidelines for the optimal choice of the technique's four parameters (POS filter, text corpus for semantic weights, term-to-term measure, vector-to-vector measure):

   - *POS filter*: in lexical definitions, nouns (*NN*) convey most semantic similarity (ρ =.76 $\pm$ .03 for similarity, .74 for relatedness). The addition of verbs slightly affects the results, at times for the better, at times for the worse (ρ =.75 $\pm$ .03 for similarity, .73 for relatedness). Verbs in isolation obtain very low cognitive plausibility.
   - *Corpus* C: to obtain the highest cognitive plausibility, a general text corpus (extracted from the *Irish Independent* newspaper) is preferable (.75 $\pm$ .02 for similarity, .74 for relatedness). A domain specific dataset (OSM Wiki website), and uniform weights (*Null* corpus) reach lower results ($\leq$ .69).
   - *Term-to-term measure* sim_t: Three measures appear to be the best choice: *lch* [170], .75 $\pm$ .03 for similarity, and .74 relatedness; *path* [239], .74 $\pm$ .02 for similarity, and .74 relatedness; *vector* [232], .73 $\pm$ .03 for similarity, and .72 relatedness. The other measures obtain lower results ($\leq$ .69).
   - *Vector-to-vector measure* sim_v: The vector-to-vector similarity *com* by Corley and Mihalcea [54] largely outperforms *fes* by Fernando and Stevenson [75].

7. Hybrid NetLexSiM: network and lexical components are integrated into a hybrid measure, using two combination techniques, score combination s_{sc} and ranking combination s_{rk}, with a combination factor α (Section 3.8). The empirical evidence described in Section 6.5 suggests the following points:

   - The hybrid measures, combining network-based and lexical semantic similarity, tend to outperform individual measures, obtaining higher cognitive plausibility (ρ =.84 for similarity, and .81 for relatedness).
   - Of the two combination methods introduced in Section 3.8, ranking combination s_{rk} obtains more promising results than direct score combination s_{sc}.
   - Both components, network-based and lexical, contribute to increasing the cognitive plausibility of the hybrid measures. With optimal

combination factors ($\alpha \in [.4, .6]$), hybrid measures outperform individual measures in the vast majority of the cases under consideration ($> 89\%$).

8. The similarity jury: in the context of lexical similarity for OSM concepts, the combination of similarity measures can obtain better cognitive plausibility than individual measures. This approach is analogous to the combination of disagreeing expert opinions (Section 5.7). Empirical evidence grounds the following points:
   - A similarity jury is generally more cognitively plausible than its members in isolation (partial success ratio $> 84.6\%$). By contrast, the jury is generally less cognitively plausible than the best of its members (total success ratio $< 46.1\%$).
   - In a context of limited information in which the optimal measure is unknown, it is reasonable to rely on a jury rather than on an arbitrary measure.
   - The similarity jury is consistent with the fact that, as Cooke and Goossens [53] pointed out, "a group of experts tends to perform better than the average solitary expert, but the best individual in the group often outperforms the group as a whole" (p. 644).
9. A holistic, viewport-based GIR: this system treats map viewports as documents, and extracts semantic descriptors. The descriptors are based on a Vector Space Model (VSM), and can be indexed, clustered, and easily searched. Thus, the user can retrieve map viewports that are semantically similar to a given query viewport (Section 3.9).

Throughout this thesis, these aspects of our contribution were framed in current research areas, outlined, discussed, and evaluated. Part of this work has been published in international outlets [19, 20, 21, 22, 23, 24, 26, 200]. The next section details the limitations of our contribution to the area of OSM semantics.

7.3 Limitations

This section has the purpose of discussing the limitations of the work presented in this thesis, pointing out possible solutions to overcome them. Our efforts to provide semantic support for grassroots project OSM consist of the OSM Semantic Network, and NetLexSiM, a novel approach to computing semantic similarity of OSM concepts. A set of guidelines to obtain optimal results from NetLexSiM are provided. The remainder of this section discusses the issues we have identified while conducting this body of research, for each aspect of our contributions summarised in the previous section.

In Section 3.4 we outlined a distinction between strong and weak semantics, pointing out that our work aims at increasing the position of OSM on

the semantic spectrum, i.e. going from weak to strong semantics. Although the work presented in this thesis provides semantic support for OSM, it is important to acknowledge and discuss the limitations of each aspect of our contribution, covering in particular the OSM Semantic Network, GeReSiD, and NetLexSiM:

OSM Semantic Network

- This structure is a lightweight semantic network, not a full-fledged ontology. It contains concept definitions expressed in natural language, and relationships between the concepts, such as implication, combination, and general relatedness (i.e. unqualified hyperlinks). Moreover, the network does *not* provide either a taxonomic hierarchy of concepts, or formal definitions of concept parts, functions, attributes, or cardinality constraints for the concepts. In this sense, the network is a 'weak' geo-semantic resource (see Section 3.4).
- The network is extracted automatically from wiki text, generated by OSM contributors. As it is the case with crowdsourced resources, the presence of noise is an important issue. Noise is present in the OSM Semantic Network mostly in the form of long and ambiguous definitions, and in unclear links between concepts. Techniques of noise removal, based for example on machine learning, might help automatically remove sections of the lexical definitions whose semantics is redundant and/or unclear.
- The network only includes concepts expressed in English, and their definitions in other European languages. However, OSM is a highly multi-cultural project, including more than 200 sub-projects geared towards specific geographic areas.[6]

Geo Relatedness and Similarity Dataset (GeReSiD)

- The lack of a narrow *context* in the online survey for the relatedness/similarity assessment is certainly the main limitation of the gold standard [150]. The subjects decided autonomously the relevant criteria for the assessment.
- The interrater reliability (IRR) and interrater agreement (IRA) of the dataset ($\approx .65$) are comparable to similar datasets, such as that by Rodríguez and Egenhofer [250]. However, there are no hard criteria to assess the quality of such datasets, so the validation process remains rather arbitrary.
- This survey focused on concepts from the English-speaking OSM. Other major world languages, such as Chinese and Spanish, should be included in the study to investigate the cross-cultural generalisability of such psychological tests.

[6] `http://wiki.openstreetmap.org/wiki/List_of_territory_based_projects`

Network-Lexical Similarity Measure (NetLexSiM)

- This approach to computing semantic similarity of OSM concepts does not model the *context* of the similarity assessment [150]. The measure cannot discriminate between different contexts, such as similarity of *structure* as opposed to *affordance* of man-made features.
- NetLexSiM does not include spatial aspects of the geographic concepts, such as typical size, shape, spatial distribution, etc. The measure focuses uniquely on the intensional content of the OSM Semantic Network, without considering instances of such concepts in OSM.
- This approach focuses exclusively on the conceptual level, and more work is needed to integrate NetLexSiM into an instance-to-instance similarity measure. The viewport-based holistic similarity constitutes a suitable test bed for such an integration. NetLexSiM would provide a fine-grained concept-to-concept similarity measure that could inform the viewport-to-viewport approach, obtaining higher cognitive plausibility.

Network NetLexSiM

- In NetLexSiM, we focused on co-citation measures, obtaining satisfactory results. However, other promising graph-theoretical similarity measures exist, such as random walks [240].
- The co-citation measures perform considerably better for semantic similarity than for relatedness. Specific measures for semantic relatedness should be included in NetLexSiM.

Lexical NetLexSiM

- Our approach is based on a bag-of-words (BOW) model, losing syntactic aspects of the lexical definitions. For example, the sentences "X is smaller than Y" and "Y is smaller than X" result in the same semantic representation. The detection of linguistic patterns in the lexical definitions might be used to detect specific semantic relations in the raw text.
- The term-to-term measures rely on WordNet. WordNet has been widely used in natural language processing, but has well-known limitations [87]. Above all, it has limited coverage of words/concepts. As it is manually updated by a small number of experts, the coverage of the network is severely limited. The number of nouns modelled in WordNet is about 80,000, while crowd-sourced knowledge bases such as DBpedia or YAGO contain millions of entities, although their quality tends to be more variable than that of WordNet.
- The coverage of geographic concepts in WordNet is limited, and GeoWordNet may provide a richer source of semantic similarity

in the geospatial domain [92]. Although some corpus-based information is included in the approach (i.e. the semantic weights of vectors), NetLexSiM is predominantly knowledge-based. Corpus-based measures should be taken into account [241].

- The lexical component of NetLexSiM approximates semantic relatedness and similarity in a comparable way. Ideally, the measure should be able to work in separate 'modes,' i.e. geared towards semantic relatedness *or* similarity.
- The temporal complexity of the approach is cubic. In particular, the WordNet-based similarity approaches constitute a bottleneck, and need pre-computation (see Section 5.6).

Holistic, viewport-based GIR system

- In this system, the map viewports are described through semantic descriptors, i.e. vectors built on the feature types contained on the viewport. Hence, the effectiveness of such viewport-based depends on the quality, expressivity and accuracy of such descriptors. As similarity judgements are highly context-sensitive, the system should be able to adapt to different user needs, weighting the descriptors dynamically.
- The weighting schemes that we considered are highly sensitive to the number of features in the viewport. However, descriptors based on simple feature count are unlikely to capture the complexity of viewport semantics.
- Geostatistical indices can capture complex topological patterns in spatial data, such as fragmentation, spatial heterogeneity, spatial entropy, diversity, and spatial autocorrelation. Descriptors based on these indices will help develop cognitively plausible viewport-based similarity techniques beyond this proof-of-concept.

While some of these limitations have a manageable solution (e.g. the multi-language extension of the OSM Semantic Network), others present deeper challenges, approachable via larger research efforts, such as the issue of context in geo-semantic similarity. Our contributions to the semantics of OSM should be considered as promising starting points to provide 'stronger' semantics in real-world applications, enabling a more effective usage of the informational wealth generated within VGI. The next section outlines our plans for future work, pointing out directions that we deem promising.

7.4 Conclusions and future work

In this thesis, we have outlined several contributions to the topic of geo-semantics in the context of spatial crowdsourcing and Volunteered Geographic Information (VGI). This section will draw conclusions on this body of work,

discussing how our work advances the research area, and will outline research directions to extend our contribution further. Finally, we will consider the general perspectives of VGI within the so-called Digital Earth.

The body of work presented in this thesis has aimed at providing semantic support for crowdsourced geographic data. The broad challenge we focused on was that of the semantic gap within VGI, i.e. the semantic ambiguity that hinders the effective usage of the data in complex automated tasks. When using data generated through open collaborative processes, users face several semantic challenges, mainly caused by the lack of a centralised, expert-authored ontological ground. Focusing on the most prominent instance of VGI, OpenStreetMap (OSM), we aimed at capturing its rich and dynamic geographic conceptualisation in a formal knowledge base.

As a result, we developed a novel resource, the OSM Semantic Network, designed to facilitate the usage of the vast and ever-growing OSM vector dataset. Unlike similar semantic projects, this machine-readable network represents the rich conceptualisation defined by contributors on the OSM Wiki website, offering a valuable tool to assist general users in information integration. The detailed description of the OSM geographic concepts provides a basis for automated alignment techniques between OSM and heterogenous data sources, alleviating the pervasive issue of semantic ambiguity. Analogously, the OSM Semantic Network can support the usage of user-generated data in Geographic Information Retrieval (GIR), helping the identification of relevant geographic concepts in fine-grained automatic semantic processing.

Subsequently, by exploiting the geographic knowledge encoded in the OSM Semantic Network, we devised a hybrid semantic similarity measure for geographic concepts, the Network-Lexical Similarity Measure (NetLexSiM), based on two components (network-based and lexical similarity), obtaining high cognitive plausibility. This contribution offers useful results both from a pragmatic and a theoretical viewpoint. From a pragmatic perspective, the evaluation of NetLexSiM provides general users with precise guidelines to compute the semantic similarity of crowdsourced concepts, via cognitively plausible techniques. The benefits derive from the crucial importance of the computation of semantic similarity, which is necessary to tackle a variety of advanced problems in data mining, GIR, natural language processing, and information integration.

From a theoretical viewpoint, our Geo Relatedness and Similarity Dataset (GeReSiD) provides researchers with a novel gold standard. Unlike other gold standards, GeReSiD explicitly captures the difference between general semantic *relatedness* and the more specific semantic *similarity*, enabling a direct, empirical comparison between the two. Such a dataset can help clarify the terminological confusion in this research area, enabling researchers to compare and contrast the effectiveness of any similarity or relatedness measure within the geographic domain. Advancing this research area may have positive outcomes for a number of advanced real-world geo-applications, for which semantic inter-operability and processing are key.

Although we focus on OSM, the approach we adopted for the OSM Se-

mantic Network and NetLexSiM can be applied to any other crowdsourced dataset, on which concepts are defined informally on a wiki website. These contributions constitute a semantic support for VGI, and can be extended in several promising directions. Among many other possible directions for future research, we identify the following as particularly important:

- The OSM Semantic Network should become a multi-lingual resource, tapping the large non-English sections of the OSM Wiki website (Section 3.3). The OSM Wiki Crawler would easily accommodate such an improvement.
- The NetLexSiM measure is largely knowledge-based (Section 3.6). Corpus-based techniques, such as Latent Semantic Analysis (LSA), should be included in the measure [241]. GeoWordNet may replace WordNet as a source of knowledge for term semantic similarity [92]. Specific measures for semantic relatedness should be devised. Such measures of semantic relatedness and similarity can be seen as a module to be plugged into any application using OSM geographic data, supporting its thin semantic model for GIR, map personalisation, information integration, and other semantic-centered tasks.
- NetLexSiM focuses on OSM classes, and not instances. Ultimately, the concept-to-concept similarity may be used to support a feature-to-feature similarity/relatedness measure, including instance-based similarity measures [215]. In a feature-to-feature similarity measure, spatial aspects of similarity, such as area, shape, proximity, should be given a strong emphasis.
- Our approach to combining similarity measures, the 'similarity jury,' provides a more reliable way to compute semantic similarity in a context in which the optimal measure is unknown (Section 5.7). However, the evaluation of the similarity jury was conducted only on the lexical definitions of OSM concepts. A more extensive evaluation could be conducted on non-spatial standards, combining the results through meta-analytical methods (see Section 2.5.5 for a survey of similarity gold standards).
- Our holistic, viewport-based semantic system represents an alternative approach to GIR (Section 3.9). A preliminary evaluation was presented. A larger implementation, e.g. including a subset of the OSM world vector map, would enable a further exploration and evaluation of this novel holistic approach, highlighting the strengths and weaknesses of the viewport semantic descriptors. Our similarity measure NetLexSiM can be integrated into this GIR system, heading towards a framework to compute similarity of places.

These research directions are inscribed in the large, interdisciplinary framework of what has been dubbed 'Digital Earth' by Al Gore [101, 98]. The construction of a virtual representation Earth, as Graham [102] aptly put it,

is "a worldwide engineering project that is unprecedented in scale or scope" (p. 422). This project is not a top-down, centralised enterprise, but involves a number of diverse individual and collective actors, in a complex system of inter-locking feedback loops. While the technological infrastructure for a virtual planet was developed mostly within expert-dominated public institutions, a relevant part of its content is now being produced by volunteers through Web 2.0 paradigm [96].

As a result, Volunteered Geographic Information (VGI) is not produced following strict scientific and technical standards, but through a collaborative negotiation of volunteers of varying levels of expertise, which poses specific challenges. The crowdsourcing approach has evident limitations, including recurring vandalism, variable spatial/semantic quality, the rural/urban divide, and the economic divide, which cause practical challenges for the project contributors and users. For this reason, VGI was at first received with benign scepticism within the Geographic Information Systems (GIS) community, mainly because of the lack of precise metrics to assess its data quality and reliability [32, 212].

However, the relentless growth of VGI data and contributors suggests that the driving forces of the project – emphasis on free data, open semantics, and local mapping – largely overcome its drawbacks [223]. Despite its flaws, VGI has experienced extraordinary popular success, and has quickly been recognised as a powerful resource for gathering geographic knowledge, reducing mapping costs, and supporting profit and non-profit processes [126, 105]. Amateur users can generate knowledge about their surroundings: as Goodchild [97] put it, "we are all experts in our own local communities" (p. 95).

All these trends are clearly visible in the leading VGI project that we have focused on in this thesis, the grassroots OpenStreetMap (OSM). The history of OSM bears, *mutatis mutandis*, striking parallels with Wikipedia, which, despite its fierce critics, has become a central resource for millions of Internet users in less than a decade [81]. As recent developments in the Web mapping market suggest [223, 115], the importance of spatial crowdsourcing projects such as OSM is likely to keep growing in the near future: VGI, in short, is here to stay.

One of the key issues of the collaborative side of the Digital Earth lies, indeed, in the models adopted to create and utilise a bottom-up, shared ground for geographic meaning [205]. On the OSM Wiki website, for example, users define concepts that describe the actual vector map. In the area of geo-semantics, we distinguish between two paradigms to ground geographic meaning in digital information: at one end, 'weak geo-semantics' rely on simple schemas and Web 2.0 models such as collaborative tagging, while 'strong geo-semantics' utilises top-down formalised geo-ontologies, inferential engines originated in the context of the Semantic Web and artificial intelligence (AI). Although the idea of an *actual* 'strong semantics' is hindered by the limitations of AI, the Semantic Web efforts have started to deliver notable outcomes not only in academic research, but also in profit-driven businesses [42].

In this thesis, we aimed at making the OSM semantics 'stronger,' providing semantic support for its tagging model. However, this semantic gap

between 'strong' and 'weak' approaches to geo-semantics is still large, and constitutes an interdisciplinary key challenge [121, 104]. The OSM semantic model, based on open collaborative tagging, cannot be constrained in a full-fledged ontology, which requires a high degree of authoritarian centralisation. Hence, a critical direction for future geo-semantic research consists of developing further techniques to push the boundaries of VGI towards stronger semantics, compatibly with its collaborative openness. Although the convergence of all available geographic information towards what Al Gore called a 'geoportal' is likely to remain an unfulfilled vision, scientific advances in geo-information integration certainly rest upon our understanding of data semantics, and the development of novel approaches to metadata [98].

Further research to fill this gap offers exciting perspectives both at the theoretical level – studying how geographic meaning is grounded across open virtual communities – and at the pragmatic level, enabling a higher degree of semantic support for countless GeoWeb applications. Rich, machine-readable descriptions of geographic concepts and entities might not lead to something as powerful as the original vision of the Semantic Web, but certainly contribute to engineering effective solutions to map personalisation, information integration, data mining, and GIR [257, 139, 160]. Another valuable contribution might consist of metrics for *semantic quality*, an aspect complementary to spatial quality, assisting OSM contributors in detecting and solving semantic criticalities [212].

The economic, social, epistemological and technological consequences of the Digital Earth and VGI are far-reaching and hard to predict – or to overestimate. As Graham [102] argued, new virtual layers of geographic representations are constantly being expanded and moulded in a dialectic relationship between physical and digital spaces. Understanding how the semantics of such layers is structured and how it can be formalised and grounded for machines is a fundamental pre-condition to increase what we might call the 'geographic intelligence' of collaborative, crowdsourced neogeographic applications.

REFERENCES

1. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27), ACL.

2. Ahlqvist, O. (2004). A parameterized representation of uncertain conceptual spaces. *Transactions in GIS*, *8*(4), 493–514.

3. Ahlqvist, O. (2005). Using semantic similarity metrics to uncover category and land cover change. In: *GeoSpatial Semantics* (pp. 107–119), Springer, LNCS, vol. 3799.

4. Ahlqvist, O. (2011). On the (limited) difference between feature and geometric semantic similarity models. In: *GeoSpatial Semantics* (pp. 124–132), Springer, LNCS, vol. 6631.

5. Albadvi, A., Shahbazi, M. (2009). A hybrid recommendation technique based on product category attributes. *Expert Systems with Applications*, *36*(9), 11,480–11,488.

6. Albanese, M., Picariello, A., Sansone, C., Sansone, L. (2004). Web personalization based on static information and dynamic user behavior. In: *WIDM '04: Proceedings of the 6th annual ACM International Workshop on Web Information and Data Management* (pp. 80–87), ACM.

7. Odon de Alencar, R., Davis Jr, C., Gonçalves, M. (2010). Geographical classification of documents using evidence from Wikipedia. In: *Proceedings of the 6th ACM Workshop on Geographic Information Retrieval, GIR'10* (pp. 12:1–12:8), ACM.

8. Altman, D., Gardner, M. (1988). Statistics in medicine: Calculating confidence intervals for regression and correlation. *British Medical Journal (Clinical research ed)*, *296*(6631), 1238–1242.

9. Amsler, R. (1972). Applications of citation-based automatic classification. Technical Report 14, Linguistics Research Center, Austin, TX.

10. Andreasen, M., Nielsen, H., Schrøder, S., Stage, J. (2006). Usability in open source software development: opinions and practice. *Information technology and control*, *25*(3A), 303–12.

11. Ankolekar, A., Krötzsch, M., Tran, T., Vrandecic, D. (2008). The two cultures: Mashing up Web 2.0 and the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(1), 70–75.

12. Antrop, M., Van Eetvelde, V. (2000). Holistic aspects of suburban landscapes: visual image interpretation and landscape metrics. *Landscape and Urban Planning*, *50*(1-3), 43–58.

13. Aristotle (1931 (350 BC)). On Memory and Reminiscence. In: *The Works of Aristotle*, URL `http://classics.mit.edu/Aristotle/memory.html`, trans. J. I. Beare.

14. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In: *The Semantic Web* (pp. 722–735), Springer, LNCS, vol. 4825.

15. Auer, S., Lehmann, J., Hellmann, S. (2009). LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In: *Proceedings of the International Semantic Web Conference, ISWC 09* (pp. 731–746), Springer, LNCS, vol. 5823.

16. Baader, F., Horrocks, I., Sattler, U. (2005). Description Logics as Ontology Languages for the Semantic Web. In: *Mechanizing Mathematical Reasoning* (pp. 228–248), Springer, LNCS, vol. 2605.

17. Bakillah, M., Bédard, Y., Mostafavi, M., Brodeur, J. (2009). SIM-NET: A View-Based Semantic Similarity Model for Ad Hoc Networks of Geospatial Databases. *Transactions in GIS*, *13*(5-6), 417–447.

18. Balabanovic, M., Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Communications of the ACM*, *40*, 66–72.

19. Ballatore, A., Bertolotto, M. (2011). Semantically Enriching VGI in Support of Implicit Feedback Analysis. In: *Proceedings of the Web and Wireless Geographical Information Systems International Symposium (W2GIS 2011)*, LNCS, vol. 6574, Springer, pp. 78–93.

20. Ballatore, A., McArdle, G., Kelly, C., Bertolotto, M. (2010). RecoMap: An Interactive and Adaptive Map-based Recommender. In: *Proceedings of the 2010 ACM Symposium on Applied Computing, SAC '10* (pp. 887–891), ACM.

21. Ballatore, A., McArdle, G., Tahir, A., Bertolotto, M. (2011). A Comparison of Open Source Geospatial Technologies for Web Mapping. *International Journal of Web Engineering and Technology*, *6*(4), 354–374.

22. Ballatore, A., Bertolotto, M., Wilson, D. (2012). Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap. *Knowledge and Information Systems*, pp. 1–21.

23. Ballatore, A., Wilson, D., Bertolotto, M. (2012). A Holistic Semantic Similarity Measure for Viewports in Interactive Maps. In: *Proceedings of the Web and Wireless Geographical Information Systems International Symposium (W2GIS 2012)*, LNCS, vol. 7236, Springer, pp. 151–166.

24. Ballatore, A., Wilson, D., Bertolotto, M. (2012). The Similarity Jury: Combining expert judgements on geographic concepts. In: *Advances in Conceptual Modeling. ER 2012 Workshops (SeCoGIS)*, LNCS, vol. 7518, Springer, pp. 231–240.

25. Ballatore, A., Bertolotto, M., Wilson, D. (2013). Grounding Linked Open Data in WordNet: The Case of the OSM Semantic Network. In: *Proceedings of the Web and Wireless Geographical Information Systems International Symposium (W2GIS 2013)*, LNCS, vol. 7820, Springer, pp. 1–15.

26. Ballatore, A., Wilson, D., Bertolotto, M. (2013). A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web. In: Pasi, G., Bordogna, G., Jain, L. (Eds) *Quality Issues in the Management of Web Information*, Intelligent Systems Reference Library, vol. 50, Springer, pp. 93–120.

27. Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D. (1999). Beyond Kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*(1), 3–23.

28. Banerjee, S., Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In: *Computational Linguistics and Intelligent Text Processing* (pp. 117–171), Springer, LNCS, vol. 2276.

29. Barak, A., English, N. (2002). Prospects and limitations of psychological testing on the Internet. *Journal of Technology in Human Services*, *19*(2-3), 65–89.

30. Bennett, J. (2010). *OpenStreetMap*. Birmingham, UK: Packt.

31. Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 28–37.

32. Bishr, M., Kuhn, W. (2007). Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In: *The European Information Society: Leading the Way with Geo-information, Proceedings of the 10th AGILE Conference* (pp. 365–387), Springer, LNGC.

33. Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22.

34. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S. (2009). DBpedia – A crystallization point for the Web of Data. *Journal of Web Semantics*, *7*(3), 154–165.

35. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (pp. 1247–1250), ACM.

36. Bos, J. (2011). A Survey of Computational Semantics: Representation, Inference and Knowledge in Wide-Coverage Text Understanding. *Language and Linguistics Compass*, *5*(6), 336–366.

37. Brunato, M., Battiti, R. (2003). PILGRIM: A Location Broker and Mobility-Aware Recommendation System. In: *IEEE International Conference on Pervasive Computing and Communications* (pp. 265–272), IEEE.

38. Budanitsky, A., Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: *Workshop on WordNet and Other Lexical Resources, Second Meeeting of the North American Chapter of the Association for Computational Linguistics* (pp. 29–34), ACL.

39. Budanitsky, A., Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, *32*(1), 13–47.

40. Budescu, D., Rantilla, A. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, *104*(3), 371–398.

41. Burgess, C., Lund, K. (1995). Hyperspace Analog to Language (HAL): A General Model of Semantic Representation. In: *Proceedings of the 36th Annual Meeting of the Psychonomic Society* (pp. 177–210), Psychonomic Society, vol. 12.

42. Cardoso, J. (2007). The Semantic Web Vision: Where are we? *Intelligent Systems*, *22*(5), 84–88.

43. Carston, R. (2008). Linguistic communication and the semantics/pragmatics distinction. *Synthese*, *165*(3), 321–345.

44. Cazzanti, L., Gupta, M. (2006). Information-theoretic and set-theoretic similarity. In: *Proceedings of the 2006 IEEE International Symposium on Information Theory* (pp. 1836–1840), IEEE.

45. Chester, M. (2003). Multiple Measures and High-Stakes Decisions: A Framework for Combining Measures. *Educational Measurement: Issues and Practice*, *22*(2), 32–41.

46. Chomsky, N. (2002 (1957)). *Syntactic structures*, 2nd edn. Berlin: Walter de Gruyter.

47. Cimiano, P., Völker, J. (2005). Towards large-scale, open-domain and ontology-based named entity classification. In: *Recent Advances in Natural Language Processing, RANLP 2005* (pp. 166–172), ACL.

48. Clemen, R., Winkler, R. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*(2), 187–203.

49. Coast, S. (2010). OpenStreetMap - The Best Map (February 19, 2010). *OpenGeoData*, URL `http://opengeodata.org/openstreetmap-the-best-map`.

50. Codescu, M., Horsinka, G., Kutz, O., Mossakowski, T., Rau, R. (2011). OSMonto - An Ontology of OpenStreetMap Tags. URL `http://wiki.openstreetmap.org/wiki/OSMonto`.

51. Coleman, D., Georgiadou, Y., Labonte, J. (2009). Volunteered Geographic Information: the nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, *4*(2009), 332–358.

52. Collins, A., Loftus, E. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, *82*(6), 407–428.

53. Cooke, R., Goossens, L. (2004). Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research*, *7*(6), 643–656.

54. Corley, C., Mihalcea, R. (2005). Measuring the semantic similarity of texts. In: *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment* (pp. 13–18), ACL.

55. Corver, N., van Riemsdijk, H. (2001). *Semi-lexical categories: the function of content words and the content of function words*. Berlin: Walter de Gruyter.

56. Cowart, M. (2012). Embodied Cognition. *The Internet Encyclopedia of Philosophy*, URL `http://www.iep.utm.edu/embodcog`.

57. Cureton, E. (1958). The average Spearman rank criterion correlation when ties are present. *Psychometrika*, *23*(3), 271–272.

58. Dawes, J. (2008). Do data characteristics change according to the number of scale points used? *International Journal of Market Research*, *50*(1), 61–78.

59. Deshpande, A., Riehle, D. (2008). The total growth of open source. In: *Open Source Development, Communities and Quality* (pp. 197–209), Springer, IFIP Advances in Information and Communication Technology, vol. 275.

60. Diener, M., Hilsenroth, M., Weinberger, J. (2009). A primer on meta-analysis of correlation coefficients: The relationship between patient-reported therapeutic alliance and adult attachment style as an illustration. *Psychotherapy Research*, *4-5*(19), 519–526.

61. Ding, L., Kolari, P., Ding, Z., Avancha, S. (2007). Using ontologies in the Semantic Web: A survey. In: *Ontologies* (pp. 79–113), Springer, Integrated Series in Information Systems, vol. 14.

62. Doctorow, C. (2001). Metacrap: Putting the torch to seven straw-men of the meta-utopia. URL `http://www.well.com/~doctorow/metacrap.htm`.

63. Dolan, B., Quirk, C., Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: *Proceedings of the 20th international conference on Computational Linguistics* (pp. 350–357), ACL.

64. Efrati, A. (March 15, 2012). Google Gives Search a Refresh. *Wall Street Journal*, URL `http://online.wsj.com/article/SB10001424052702304459804577281842851136290.html`.

65. Egenhofer, M. (2002). Toward the Semantic Geospatial Web. In: *Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems* (pp. 1–4), ACM.

66. Egenhofer, M., Mark, D. (1995). Naive geography. In: Frank, A. U., Kuhn, W. (Eds) *Spatial Information Theory: A Theoretical Basis for GIS* (pp. 1–15), Springer, LNCS, vol. 1329.

67. Elwood, S., Goodchild, M., Sui, D. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, *102*(3), 571–590.

68. Euzenat, J., Ferrara, A., Meilicke, C., Nikolov, A., Pane, J., Scharffe, F., Shvaiko, P., et al. (2010). Results of the Ontology Alignment Evaluation Initiative 2010. Technical report, Ontology Alignment Evaluation Initiative.

69. Evans, J., Mathur, A. (2005). The value of online surveys. *Internet Research*, *15*(2), 195–219.

70. Fabrikant, S., Ruocco, M., Middleton, R., Montello, D., Jörgensen, C. (2002). The First Law of Cognitive Geography: Distance and Similarity in Semantic Space. In: *Proceedings of the Second International Conference on Geographic Information Science (GIScience 2002)* (pp. 31–33), Springer, LNCS, vol. 2478.

71. Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E. (2004). Comparing and aggregating rankings with ties. In: *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '04* (pp. 47–58), ACM.

72. Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E. (2007). Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, *20*(3), 628–648.

73. Fegeas, R., Cascio, J., Lazar, R. (1992). An overview of FIPS 173, the Spatial Data Transfer Standard. *Cartography and Geographic Information Science*, *19*(5), 278–293.

74. Fellbaum, C. (ed.) (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

75. Fernando, S., Stevenson, M. (2008). A semantic similarity approach to paraphrase detection. In: *Proceedings of Computational Linguistics UK (CLUK 2008), 11th Annual Research Colloquium* (pp. 1–7), Computational Linguistics UK: .

76. Ferrell, W. (1985). Combining individual judgments. In: Wright, G. (ed.) *Behavioral decision making*, New York: Plenum Press, pp. 111–145.

77. Field, A. (2001). Meta-analysis of correlation coefficients: A monte carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*(2), 161–180.

78. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E. (2002). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131.

79. Finn, R. (1970). A Note on Estimating the Reliability of Categorical Data. *Educational and Psychological Measurement*, *30*(1), 71–76.

80. Firth, J. (1957). A synopsis of linguistic theory 1930-55. In: *Studies in Linguistic Analysis (Special Volume of the Philological Society)*, 11, London: Longman, pp. 1–31.

81. Fletcher, J. (August 18, 2009). A brief history of Wikipedia. *Time Business*, URL `http://www.time.com/time/business/article/0,8599,1917002,00.html`.

82. Floridi, L. (2009). Web 2.0 vs. the Semantic Web: A Philosophical Assessment. *Episteme*, *6*(1), 25–37.

83. Foo, N., Garner, B., Rao, A., Tsui, E. (1992). Semantic distance in conceptual graphs. In: Nagle, T., Nagle, J., Gerholz, L., Eklund, P. (Eds) *Conceptual Structures: Current Research and Practice*, New York: Ellis Horwood, pp. 149–154.

84. Frege, G. (1956). The thought: A logical inquiry. *Mind*, *65*(259), 289–311.

85. Fu, G., Jones, C., Abdelmoty, A. (2005). Ontology-based spatial query expansion in information retrieval. In: *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE* (pp. 1466–1482), Springer, LNCS, vol. 3761.

86. Gahegan, M., Agrawal, R., Jaiswal, A., Luo, J., Soon, K. (2008). A platform for visualizing and experimenting with measures of semantic similarity in ontologies and concept maps. *Transactions in GIS*, *12*(6), 713–732.

87. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. (2003). Sweetening WordNet with DOLCE. *AI magazine*, *24*(3), 13.

88. Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

89. Gartner, G., Bennett, D., Morita, T. (2007). Towards Ubiquitous Cartography. *Cartography and Geographic Information Science*, *34*(4), 247–257.

90. Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford Linguistics, New York: Oxford University Press.

91. Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., Petras, V. (2006). GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview. In: *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005* (pp. 908–919), Springer, LNCS, vol. 4022.

92. Giunchiglia, F., Maltese, V., Farazi, F., Dutta, B. (2010). GeoWordNet: A Resource for Geo-Spatial Applications. In: *The Semantic Web: Research and Applications, ESWC 2010* (pp. 121–136), Springer, LNCS, vol. 6088.

93. Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher*, *5*(10), 3–8.

94. Goldstone, R. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*(2), 125–157.

95. Goldstone, R., Son, J. (2005). Similarity. In: Holyoak, K., Morrison, R. (Eds) *Cambridge Handbook of Thinking and Reasoning*, New York: Cambridge University Press, pp. 13–36.

96. Goodchild, M. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, *69*(4), 211–221.

97. Goodchild, M. (2009). NeoGeography and the nature of geographic expertise. *Journal of Location Based Services*, *3*(2), 82–96.

98. Goodchild, M. (2012). The future of Digital Earth. *Annals of GIS*, *18*(2), 93–98.

99. Goodman, L., Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*(268), 732–764.

100. Goodman, N. (1972). Seven strictures on similarity. In: *Problems and projects*, New York: Bobbs-Merrill, pp. 437–447.

101. Gore, A. (1998). The Digital Earth: Understanding our planet in the 21st century. *The Australian Surveyor*, *43*(2), 89–91.

102. Graham, M. (2010). Neogeography and the Palimpsests of Place: Web 2.0 and the Construction of a Virtual Earth. *Tijdschrift voor economische en sociale geografie*, *101*(4), 422–436.

103. Grassi, M., Piazza, F. (2011). Towards an RDF encoding of ConceptNet. In: *Advances in Neural Networks: 8th International Symposium on Neural Networks, ISNN 2011* (pp. 558–565), Springer, LNCS, vol. 6675.

104. Greaves, M., Mika, P. (2012). Semantic Web & Web 2.0. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(1).

105. Greengard, S. (2011). Following the crowd. *Communications of the ACM*, *54*(2), 20–22.

106. Griffiths, T., Steyvers, M. (2002). Prediction and semantic association. In: Becker, S., Thrun, S., Obermayer, K. (Eds) *Advances in Neural Information Processing Systems 15*, Cambridge, MA: MIT Press, pp. 11–18.

107. Griffiths, T., Steyvers, M., Tenenbaum, J. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

108. Gruber, T. (2007). Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems*, *3*(1), 1–11.

109. Guarino, N., Giaretta, P. (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. In: Mars, N. (ed.) *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, Amsterdam, Netherlands: IOS Press, pp. 25–32.

110. Guarino, N., Oberle, D., Staab, S. (2009). What is an Ontology? In: Staab, S., Studer, R. (Eds) *Handbook on Ontologies*, 2nd edn, Springer, pp. 1–17.

111. Haav, H., Kaljuvee, A., Luts, M., Vajakas, T. (2009). Ontology-Based Retrieval of Spatially Related Objects for Location Based Services. In: *On the Move to Meaningful Internet Systems: OTM 2009* (pp. 1010–1024), Springer: , LNCS, vol. 5871.

112. Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703.

113. Haklay, M., Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, *7*(4), 12–18.

114. Haklay, M., Singleton, A., Parker, C. (2008). Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass*, *2*(6), 2011–2039.

115. Hardy, Q. (March 20, 2012). Facing Fees, Some Sites Are Bypassing Google Maps. *New York Times*, URL `http://www.nytimes.com/2012/03/20/technology/many-sites-chart-a-new-course-as-google-expands-fees.html`.

116. Harman, D. (2005). The history of IDF and its influences on IR and other fields. In: Tait, J. (ed.) *Charting a New Course: Natural Language Processing and Information Retrieval*, The Information Retrieval Series, vol. 16, Springer, pp. 69–79.

117. Harris, Z. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

118. Hars, A., Ou, S. (2002). Working for free? motivations for participating in open-source projects. *International Journal of Electronic Commerce*, *6*(3), 25–39.

119. Harzing, A. (2007). Publish or perish. URL `http://www.harzing.com/pop.htm`.

120. Havasi, C., Speer, R., Alonso, J. (2007). ConceptNet 3: A Flexible, Multilingual Semantic Network for Common Sense Knowledge. In: *Recent Advances in Natural Language Processing, RANLP 2007* (pp. 27–29), ACL.

121. Hendler, J., Berners-Lee, T. (2010). From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. *Artificial Intelligence*, *174*(2), 156–161.

122. Hill, L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In: *Research and Advanced Technology for Digital Libraries* (pp. 280–290), Springer, LNCS, vol. 1923.

123. Hiramoto, R., Sumiya, K. (2006). Web information retrieval based on user operation on digital maps. In: *Proceedings of the 14th annual ACM International Symposium on Advances in Geographic Information Systems* (pp. 99–106), ACM.

124. Hirst, G., St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (ed.) *WordNet: An electronic lexical database*, Cambridge, MA: MIT Press, pp. 305–332.

125. Hoffart, J., Suchanek, F., Berberich, K., Lewis-Kelham, E., De Melo, G., Weikum, G. (2011). YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In: *Proceedings of the 20th International Conference Companion on World Wide Web* (pp. 229–232), ACM.

126. Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. New York: Crown Business.

127. Hughes, T., Ramage, D. (2007). Lexical semantic relatedness with random graph walks. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 581–589), ACL.

128. Hume, D. (2006 (1748)). *An enquiry concerning human understanding*. Project Gutenberg, URL `http://www.gutenberg.org/ebooks/9662`.

129. Hunter, J., Schmidt, F. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: SAGE.

130. Islam, A., Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, *2*(10), 1–25.

131. James, L., Demaree, R., Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of applied psychology*, *69*(1), 85–98.

132. Janowicz, K., Keßler, C., Schwarz, M., Wilkes, M., Panov, I., Espeter, M., Bäumer, B. (2007). Algorithm, implementation and application of the SIM-DL similarity server. In: *GeoSpatial Semantics: Second International Conference, GeoS 2007* (pp. 128–145), Springer, LNCS, vol. 4853.

133. Janowicz, K., Keßler, C., Panov, I., Wilkes, M., Espeter, M., Schwarz, M. (2008). A study on the cognitive plausibility of SIM-DL similarity rankings for geographic feature types. In: Fabrikant, S., Wachowicz, M. (Eds) *The European Information Society: Taking Geoinformation Science One Step Further*, LNGC, Springer, pp. 115–134.

134. Janowicz, K., Raubal, M., Schwering, A., Kuhn, W. (2008). Semantic Similarity Measurement and Geospatial Applications. *Transactions in GIS*, *12*(6), 651–659.

135. Janowicz, K., Raubal, M., Kuhn, W. (2011). The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, *2*(1), 29–57.

136. Jeh, G., Widom, J. (2002). SimRank: A Measure of Structural-Context Similarity. In: *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD* (pp. 538–543), ACM.

137. Jiang, J., Conrath, D. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: *Proceedings of International Conference on Research in Computational Linguistics, ROCLING X* (pp. 19–33), ACL, vol. 1.

138. Jing, L., Ng, M., Huang, J. (2010). Knowledge-based vector space model for text clustering. *Knowledge and Information Systems*, *25*(1), 1–21.

139. Jones, C., Alani, H., Tudhope, D. (2001). Geographical Information Retrieval with Ontologies of Place. In: *Spatial Information Theory* (pp. 322–335), Springer, LNCS, vol. 2205.

140. Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21.

141. Jonesh, K., Walker, S., Robertson, S. (2000). A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. *Information Processing and Management*, *36*(6), 779–808.

142. Kant, I. (2003 (1781)). *The Critique of Pure Reason*. Project Gutenberg, URL `http://www.gutenberg.org/ebooks/4280`, trans. Meiklejohn, J.M.D..

143. Kaptchuk, T. (2001). The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology*, *54*(6), 541–549.

144. Katz, J. (1970). Interpretative semantics vs. generative semantics. *Foundations of Language*, *6*(2), 220–259.

145. Kavouras, M., Kokla, M. (2008). *Theories of Geographic Concepts: Ontological Approaches to Semantic Integration*. Boca Raton, FL: CRC Press.

146. Kavouras, M., Kokla, M., Tomai, E. (2005). Comparing categories among geographic ontologies. *Computers & Geosciences*, *31*(2), 145–154.

147. Kelly, D., Teevan, J. (2003). Implicit feedback for inferring user preference: a bibliography. , *37*(2), 18–28.

148. Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, *30*(1–2), 81–93.

149. Kendall, M., Smith, B. (1939). The problem of m rankings. *The annals of mathematical statistics*, *10*(3), 275–287.

150. Keßler, C. (2007). Similarity measurement in context. In: *Proceedings of the 6th International and Interdisciplinary Conference on Modeling and Using Context* (pp. 277–290), Springer, LNCS, vol. 4635.

151. Keßler, C. (2011). What is the difference? A cognitive dissimilarity measure for information retrieval result sets. *Knowledge and Information Systems*, *30*(2), 319–340.

152. Kessler, M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25.

153. Khoo, C., Na, J. (2006). Semantic relations in information science. *Annual Review of Information Science and Technology*, *40*(1), 157–207.

154. Kitayama, D., Teratani, T., Sumiya, K. (2008). Digital map restructuring method based on implicit intentions extracted from users' operations. In: *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, ICUIMC* (pp. 45–53), ACM.

155. Kitsuregawa, M., Nishida, T. (2010). Special issue on information explosion. *New Generation Computing*, *28*(3), 207–215.

156. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R. (2009). Media meets Semantic Web – How the BBC uses DBpedia and Linked Data to make connections. In: *The Semantic Web: Research and Applications* (pp. 723–737), Springer, LNCS, vol. 5554.

157. Kobsa, A. (2001). Generic User Modeling Systems. *User Modeling and User-Adapted Interaction*, *11*(1), 49–63.

158. Korpipää, P., Häkkilä, J., Kela, J., Ronkainen, S., Känsälä, I. (2004). Utilising context ontology in mobile device application personalisation. In: *Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia* (pp. 133–140), ACM.

159. Kotera, R., Kitayama, D., Oku, K., Sumiya, K. (2011). Geographical recommendation method using user's interest model based on map operation and category selection. In: *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, ICUIMC '11* (pp. 126:1–126:8), ACM.

160. Kuhn, W. (2003). Semantic reference systems. *International Journal of Geographical Information Science*, *17*(5), 405–409.

161. Kuhn, W. (2005). Geospatial Semantics: Why, of What, and How? In: *Journal of Data Semantics III. Special Issue on Semantic-based Geographical Information Systems* (pp. 1–24), Springer, LNCS, vol. 3534.

162. Kuhn, W. (2009). Semantic engineering. In: Navratil, G. (ed.) *Research Trends in Geographic Information Science*, LNGC, Springer, pp. 63–76.

163. Lakoff, G. (1988). Cognitive semantics. In: Eco, U., Santambrogio, M., Violi, P. (Eds) *Meaning and mental representations*, Bloomington, IN: Indiana University Press, pp. 119–155.

164. Lakoff, G., Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University Of Chicago Press.

165. Lakoff, G., Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Phoenix, AZ: Basic Books.

166. Landauer, T., Foltz, P., Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse processes*, *25*(2-3), 259–284.

167. Landauer, T., McNamara, D., Dennis, S., Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

168. Lanter, D., Essinger, R. (1991). User-centered graphical user interface design for GIS. Technical report, National Center for Geographic Information & Analysis (NCGIA), Santa Barbara, CA.

169. Lausen, H., Ding, Y., Stollberg, M., Fensel, D., Hernandez, R., Han, S. (2005). Semantic Web Portals: State-of-the-art Survey. *Journal of Knowledge Management*, *9*(5), 40–49.

170. Leacock, C., Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (ed.) *WordNet: An electronic lexical database*, Cambridge, MA: MIT Press, pp. 265–283.

171. LeBreton, J., Senter, J. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852.

172. Lee, M., Pincombe, B., Welsh, M. (2005). An empirical evaluation of models of text document similarity. In: *Proceedings of the XXVII Annual Conference of the Cognitive Science Society* (pp. 1254–1259), Mahwah, NJ: Lawrence Erlbaum Associates.

173. Lee, S., Kim, H., Gupta, S. (2009). Measuring open source software success. *Omega*, *37*(2), 426–438.

174. Lehrer, A. (1974). *Semantic fields and lexical structure*. Amsterdam, Netherlands: North-Holland.

175. Lehrer, A. (1985). The influence of semantic fields on semantic change. In: Fisiak, J. (ed.) *Historical Word Formation*, Berlin: Walter de Gruyter, pp. 283–296.

176. Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: *Proceedings of the 5th annual international conference on Systems documentation* (pp. 24–26), ACM.

177. Li, H., Shi, R., Chen, W., Shen, I. (2006). Image tangent space for image retrieval. *Pattern Recognition*, *2*, 1126–1130.

178. Li, P., Liu, H., Yu, J., He, J., Du, X. (2010). Fast Single-Pair SimRank Computation. In: *Proceedings of the SIAM International Conference on Data Mining, SDM2010* (pp. 571–582), Madison, WI: Omnipress.

179. Li, W., Raskin, R., Goodchild, M. (2012). Semantic similarity measurement based on knowledge mining: an artificial neural net approach. *International Journal of Geographical Information Science*, *26*(8), 1415–1435.

180. Lin, D. (1998). An information-theoretic definition of similarity. In: *Proceedings of the 15th International Conference on Machine Learning* (pp. 296–304), Morgan Kaufmann, vol. 1.

181. Lin, Z., Lyu, M., King, I. (2011). MatchSim: a novel similarity measure based on maximum neighborhood matching. *Knowledge and Information Systems*, *32*(1), 1–26.

182. Liu, J., Chen, H., Furuse, K., Kitagawa, H., Yu, J. X. (2011). On Efficient Distance-based Similarity Search. In: *Proceedings of the 11th IEEE International Conference on Data Mining Workshops* (pp. 1199–1202), IEEE.

183. Lizorkin, D., Velikhov, P., Grinev, M., Turdakov, D. (2008). Accuracy Estimate and Optimization Techniques for SimRank Computation. In: *Proceedings of the VLDB Endowment* (pp. 422–433), Very Large Data Base Endowment, vol. 1.

184. Lovász, L. (1993). Random walks on graphs: A survey. *Combinatorics*, *2*(1), 1–46.

185. Lutz, M. (2005). Ontology-based Discovery and Composition of Geographic Information Services. PhD thesis, University of Münster, Germany.

186. Lutz, M., Klien, E. (2006). Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science*, *20*(3), 233–260.

187. Mac Aoidh, E., Bertolotto, M. (2007). Improving Spatial Data Usability By Capturing User Interactions. In: *The European Information Society: Leading the Way with Geo-information, Proceedings of the 10th AGILE Conference*, LNGC, Springer, pp. 389–403.

188. Mac Aoidh, E., Bertolotto, M., Wilson, D. C. (2007). Analysis of implicit interest indicators for spatial data. In: *Proceedings of the 15th annual ACM International Symposium on Advances in Geographic Information Systems, GIS '07* (pp. 1–4), ACM.

189. Mac Aoidh, E., Bertolotto, M., Wilson, D. (2008). Understanding geospatial interests by visualizing map interaction behavior. *Information Visualization*, *7*(3), 275–286.

190. Maedche, A., Zacharias, V. (2002). Clustering ontology-based metadata in the Semantic Web. In: *Principles of Data Mining and Knowledge Discovery* (pp. 383–408), Springer, LNCS, vol. 2431.

191. Maienborn, C., von Heusinger, K., Portner, P. (2011). *Semantics: An international handbook of natural language and meaning*. Berlin: Mouton de Gruyter.

192. Mandl, T. (2011). Evaluating GIR: geography-oriented or user-oriented? *SIGSPATIAL Special*, *3*(2), 42–45.

193. Mandl, T., Gey, F., Di Nunzio, G., Ferro, N., Larson, R., Sanderson, M., Santos, D., Womser-Hacker, C., Xie, X. (2008). GeoCLEF 2007: The CLEF 2007 cross-language geographic information retrieval track overview. In: *Advances in Multilingual and Multimodal Information Retrieval* (pp. 745–772), Springer, LNCS, vol. 5152.

194. Manovich, L. (2001). *The Language of New Media*. Cambridge, MA: MIT Press.

195. Marcus, M., Marcinkiewicz, M., Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, *19*(2), 313–330.

196. Markman, A., Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*(4), 431–467.

197. Martinich, A. (1996). *The philosophy of language*. New York: Oxford University Press.

198. Martins, B., Silva, M., Andrade, L. (2005). Indexing and ranking in Geo-IR systems. In: *Proceedings of the 2005 Workshop On Geographic Information Retrieval, GIR'05* (pp. 31–34), ACM.

199. Mata, F. (2007). Geographic information retrieval by topological, geographical, and conceptual matching. In: *GeoSpatial Semantics* (pp. 98–113), Springer, LNCS, vol. 4853.

200. McArdle, G., Ballatore, A., Tahir, A., Bertolotto, M. (2010). An Open-Source Web Architecture for Adaptive Location Based Services. In: *Proceedings of the 14th International Symposium on Spatial Data Handling, SDH* (pp. 296–301), ISPRS, vol. 38.

201. Medin, D., Goldstone, R., Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, *1*(1), 64–69.

202. Middleton, S., De Roure, D., Shadbolt, N. (2001). Capturing knowledge of user preferences: ontologies in recommender systems. In: *Proceedings of the 1st International Conference on Knowledge Capture* (pp. 100–107), ACM.

203. Middleton, S., Alani, H., De Roure, D. (2002). Exploiting Synergy Between Ontologies and Recommender Systems. In: *Proceedings of the WWW2002 International Workshop on the Semantic Web* (pp. 1–10), CEUR Workshop Proceedings, vol. 55.

204. Mihalcea, R., Corley, C., Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In: *Proceedings of the Twenty-First National Conference on Artificial Intelligence* (pp. 775–780), AAAI, vol. 21.

205. Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. In: *Proceedings of the 4th International Semantic Web Conference, ISWC 2005* (pp. 522–536), Springer, LNCS, vol. 3729.

206. Miller, G., Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

207. Miller, G., Leacock, C., Tengi, R., Bunker, R. (1993). A semantic concordance. In: *Proceedings of the Workshop on Human Language Technology* (pp. 303–308), ACL.

208. Mirizzi, R., Ragone, A., Di Noia, T., Di, E. (2010). Ranking the Linked Data: The Case of DBpedia. In: *Proceedings of 10th International Conference in Web Engineering, ICWE 2010* (pp. 337–354), Springer, LNCS, vol. 6189.

209. Modrak, D. (2001). *Aristotle's theory of language and meaning*. Cambridge, UK: Cambridge University Press.

210. Mohammad, S., Hirst, G. (2012). Distributional Measures of Semantic Distance: A Survey. *Computing Research Repository (CoRR)*, *abs/1203.1858*, 1–39, URL `http://arxiv.org/abs/1203.1858`.

211. Mooney, P., Corcoran, P. (2012). Characteristics of heavily edited objects in OpenStreetMap. *Future Internet*, *4*(1), 285–305.

212. Mooney, P., Corcoran, P., Winstanley, A. (2010). Towards quality metrics for OpenStreetMap. In: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 514–517), ACM.

213. Morgan, M., Henrion, M. (1992). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. New York: Cambridge University Press.

214. Morris, J., Hirst, G. (2004). Non-classical lexical semantic relations. In: *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics* (pp. 46–51), ACL.

215. Mülligann, C., Janowicz, K., Ye, M., Lee, W. (2011). Analyzing the Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information. In: *Spatial Information Theory* (pp. 350–370), Springer, LNCS, vol. 6899.

216. Mumpower, J., Stewart, T. (1996). Expert judgement and expert disagreement. *Thinking & Reasoning*, *2*(2-3), 191–212.

217. Nadeau, D., Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

218. Nakayama, K., Hara, T., Nishio, S. (2008). Wikipedia Link Structure and Text Mining for Semantic Relation Extraction. Towards a Huge Scale Global Web Ontology. In: *Proceedings of the Workshop on Semantic Search (SemSearch 2008), 5th European Semantic Web Conference (ESWC 2008)* (pp. 59–73), CEUR Workshop Proceedings, vol. 334.

219. Naveh, Z. (2000). What is holistic landscape ecology? A conceptual introduction. *Landscape and Urban Planning*, *50*(1-3), 7–26.

220. Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, *41*(2), 10:1–10:69.

221. Nelson, D., Dyrdal, G., Goodmon, L. (2005). What is preexisting strength? Predicting free association probabilities, similarity ratings, and cued recall probabilities. *Psychonomic Bulletin & Review*, *12*(4), 711–719.

222. Nisbett, R., Peng, K., Choi, I., Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological review*, *108*(2), 291–310.

223. Nitzschke, J. (2012). OpenStreetMap's Growth Accelerates. Technical report, BeyoNav, Chicago, IL, URL http://www.beyonav.com/openstreetmaps-growth-accelerates.

224. Obrst, L. (2003). Ontologies for semantically interoperable systems. In: *Proceedings of the Twelfth International Conference on Information and Knowledge Management* (pp. 366–369), ACM.

225. Ogden, C., Richards, I. (1923). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. London: Kegan Paul.

226. O'Reilly, T. (2005). What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software. URL http://oreilly.com/web2/archive/what-is-web-20.html.

227. Overell, S. (2009). Geographic Information Retrieval: Classification, Disambiguation and Modelling. PhD thesis, Imperial College London, Department of Computing.

228. Page, L., Brin, S., Motwani, R., Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, Stanford, CA.

229. Pantel, P., Lin, D. (2002). Discovering word senses from text. In: *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD* (pp. 613–619), ACM.

230. Parker, A., Hamblen, J. (1989). Computer algorithms for plagiarism detection. *IEEE Transactions on Education*, *32*(2), 94–99.

231. Partee, M. (1972). Plato's theory of language. *Foundations of Language*, *8*(1), 113–132.

232. Patwardhan, S., Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: *Proceedings of the EACL 2006 Workshop Making Sense of Sense – Bringing Computational Linguistics and Psycholinguistics Together* (pp. 1–8), ACL, vol. 1501.

233. Pedersen, T., Kolhatkar, V. (2009). Wordnet::senserelate::allwords: a broad coverage word sense tagger that maximizes semantic relatedness. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session* (pp. 17–20), ACL.

234. Pedersen, T., Patwardhan, S., Michelizzi, J. (2004). WordNet::Similarity: measuring the relatedness of concepts. In: *Proceedings of Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session* (pp. 38–41), ACL.

235. Poo, D., Chng, B., Goh, J. (2003). A Hybrid Approach for User Profiling. In: *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, HICSS'03* (pp. 103–112), IEEE, vol. 4.

236. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P. (2010). An evaluation framework for plagiarism detection. In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 997–1005), ACL.

237. Priedhorsky, R., Terveen, L. (2008). The Computational Geowiki: What, Why, and How. In: *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW 2008* (pp. 267–276), ACM.

238. Purves, R., Jones, C. (2011). Geographic Information Retrieval. *SIGSPATIAL Special*, *3*(2), 2–4.

239. Rada, R., Mili, H., Bicknell, E., Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, *19*(1), 17–30.

240. Ramage, D., Rafferty, A., Manning, C. (2009). Random walks for text semantic similarity. In: *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing* (pp. 23–31), ACL.

241. Recchia, G., Jones, M. (2009). More data trumps smarter algorithms: comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*(3), 647–656.

242. Reichenbacher, T. (2001). Adaptive concepts for a mobile cartography. *Journal of Geographical Sciences*, *11*(1), 43–53.

243. Reimer, M. (2010). Reference. *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*, URL `http://plato.stanford.edu/archives/spr2010/entries/reference`.

244. Renda, M., Straccia, U. (2003). Web metasearch: rank vs. score based rank aggregation methods. In: *Proceedings of the 2003 ACM Symposium on Applied Computing, SAC '03* (pp. 841–846), ACM.

245. Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95* (pp. 448–453), Morgan Kaufmann, vol. 1.

246. Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520.

247. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M. (1996). Okapi at TREC-3. In: *Overview of the Third Text Retrieval Conference, TREC-3*, Diane Publishing, pp. 109–126.

248. Robinson, W. (1957). The statistical measurement of agreement. *American Sociological Review*, *22*(1), 17–25.

249. Rodgers, J., Nicewander, W. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician*, *42*(1), 59–66.

250. Rodríguez, M., Egenhofer, M. (2004). Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographical Information Science*, *18*(3), 229–256.

251. Roget, P., Kirkpatrick, B. (1998). *Roget's Thesaurus*. London: Penguin.

252. Rubenstein, H., Goodenough, J. (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, *8*(10), 627–633.

253. Russell, S., Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*, 3rd edn. Upper Saddle River, NJ: Prentice Hall.

254. Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*(5), 513–523.

255. Salton, G., Wong, A., Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

256. Sánchez, D., Batet, M., Valls, A., Gibert, K. (2010). Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems*, *35*(3), 383–413.

257. Santos, D., Chaves, M. (2006). The place of place in geographical IR. In: *3rd Workshop on Geographic Information Retrieval, SIGIR* (pp. 5–8), ACM.

258. Saussure, F. d. (2011 (1916)). *Course in general linguistics*. New York: Columbia University Press, trans. Basking W..

259. Schafer, J., Frankowski, D., Herlocker, J., Sen, S. (2007). Collaborative filtering recommender systems. In: *The Adaptive Web: Methods and Strategies of Web Personalization* (pp. 291–324), Springer, LNCS, vol. 4321.

260. Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, *24*(1), 97–123.

261. Schwering, A. (2008). Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS*, *12*(1), 5–29.

262. Schwering, A., Raubal, M. (2005). Measuring semantic similarity between geospatial conceptual regions. In: *GeoSpatial Semantics* (pp. 90–106), Springer, LNCS, vol. 3799.

263. Schwering, A., Raubal, M. (2005). Spatial relations for semantic similarity measurement. In: *Perspectives in Conceptual Modeling* (pp. 259–269), Springer, LNCS, vol. 3770.

264. Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(03), 417–424.

265. Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423 and 623–656.

266. Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

267. Shiode, N., Li, C., Batty, M., Longley, P., Maguire, D. (2004). The impact and penetration of location-based services. In: Karimi, H. A., Hammad, A. (Eds) *Telegeoinformatics: location-based computing and services*, Boca Raton, FL: CRC Press, pp. 349–366.

268. Sieg, A., Mobasher, B., Burke, R. (2007). Web Search Personalization with Ontological User Profiles. In: *Proceedings of the 16th ACM International Conference on Information and Knowledge* (pp. 525–534), ACM.

269. Siegel, S., Castellan, N. (1988). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

270. Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., Li Zhu, W. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. In: *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE* (pp. 1223–1237), Springer, LNCS, vol. 2519.

271. Small, H. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science*, *24*(4), 265–269.

272. Smith, B. (2003). Ontology. In: Floridi, L. (ed.) *The Blackwell Guide to the Philosophy of Computing and Information*, Oxford, UK: Blackwell, pp. 153–166.

273. Smith, J., Shapiro, J. (1989). The occurrence of holistic categorization. *Journal of Memory and Language*, *28*(4), 386–399.

274. Sowa, J. (1991). *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo, CA: Morgan Kaufmann.

275. Speaks, J. (2011). Theories of meaning. *The Stanford Encyclopedia of Philosophy (Summer 2011 Edition)*, URL `http://plato.stanford.edu/archives/sum2011/entries/meaning`.

276. Spearman, C. (1904). The Proof and Measurement of Association Between Two Things. *The American Journal of Psychology*, *15*(1), 72–101.

277. Stallman, R. (2009). Viewpoint: Why 'open source' misses the point of free software. *Communications of the ACM*, *52*(6), 31–33.

278. Storms, G., Van Mechelen, I., De Boeck, P. (1994). Structural analysis of the intension and extension of semantic concepts. *European Journal of Cognitive Psychology*, *6*(1), 43–75.

279. Strube, M., Ponzetto, S. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In: *Proceedings of the 21st National Conference on Artificial Intelligence, AAAI'06* (pp. 1419–1424), AAAI, vol. 2.

280. Suchanek, F., Kasneci, G., Weikum, G. (2007). YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: *Proceedings of the 16th International Conference on World Wide Web* (pp. 697–706), ACM.

281. Sui, D. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems*, *32*(1), 1–5.

282. Sumiya, K. (2009). Less-conscious information retrieval techniques for Location Based Services. In: *Proceedings of the 2009 International Workshop on Location Based Social Networks, GIS '09* (pp. 69–72), ACM.

283. Tahir, A., McArdle, G., Ballatore, A., Bertolotto, M. (2010). Collaborative Filtering: A Group Profiling Algorithm for Personalisation in a Spatial Recommender System. In: *Proceedings of Geoinformatik 2010* (pp. 44–50), Münster, Germany: IFGI Prints.

284. Tan, H., Yu, P., Li, X., Yang, Y. (2011). Digital image similarity metrics and their performances. In: *Proceedings of the 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce, AIMSEC 2011* (pp. 3922–3925), IEEE.

285. Tarski, A. (1944). The semantic conception of truth: and the foundations of semantics. *Philosophy and phenomenological research*, *4*(3), 341–376.

286. Thiagarajan, R., Manjunath, G., Stumptner, M. (2008). Computing semantic similarity using ontologies. Technical report, Hewlett-Packard Labs, Bangalore, India.

287. Tichy, W. (1998). Should computer scientists experiment more? *Computer*, *31*(5), 32–40.

288. Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. In: *Economic geography. Supplement: Proceedings. International Geographical Union. Commission on Quantitative Methods*, vol. 46, Worcester, MA: Clark University, pp. 234–240.

289. Torre, I. (2009). Adaptive systems in the era of the semantic and social web, a survey. *User Modeling and User-Adapted Interaction*, *19*(5), 433–486.

290. Toutanova, K., Klein, D., Manning, C., Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (pp. 173–180), ACL, vol. 1.

291. Turner, A. (2006). *Introduction to Neogeography*. Sebastopol, CA: O'Reilly Media.

292. Turney, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: *Proceedings of the 12th European Conference on Machine Learning, ECML'01* (pp. 491–502), Springer, LNAI, vol. 2167.

293. Turney, P. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*(3), 379–416.

294. Turney, P., Pantel, P., et al. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141–188.

295. Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327–352.

296. Uebersax, J. (2006). Likert scales: dispelling the confusion. *Statistical Methods for Rater Agreement website*, URL `http://john-uebersax.com/stat/likert.htm`.

297. Vander Wal, T. (2007). Folksonomy. URL `http://vanderwal.net/folksonomy.html`.

298. Varzi, A. (2001). Vagueness in geography. *Philosophy & Geography*, *4*(1), 49–65.

299. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R. (2006). Semantic Wikipedia. In: *Proceedings of the 15th International Conference on World Wide Web* (pp. 585–594), ACM.

300. Vosniadou, S., Ortony, A. (1989). Similarity and analogical reasoning: a synthesis. In: Vosniadou, S., Ortony, A. (Eds) *Similarity and analogical reasoning*, Cambridge, UK: Cambridge University Press, pp. 1–19.

301. Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval* (pp. 5–7), Springer, LNCS, vol. 2069.

302. Wainer, J., Novoa Barsottini, C., Lacerda, D., Magalhães de Marco, L. (2009). Empirical evaluation in Computer Science research published by ACM. *Information and Software Technology*, *51*(6), 1081–1085.

303. Wan, X. (2008). Beyond Topical Similarity: A Structural Similarity Measure for Retrieving Highly Similar Documents. *Knowledge and Information Systems*, *15*(1), 55–73.

304. Wang, T., Hirst, G. (2011). Refining the notions of depth and density in WordNet-based semantic similarity measures. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1003–1011), ACL.

305. Weakliam, J., Bertolotto, M., Wilson, D. (2005). Implicit interaction profiling for recommending spatial content. In: *Proceedings of the 13th annual ACM International Workshop on Geographic Information Systems, GIS '05* (pp. 285–294), ACM.

306. Weibelzahl, S., Weber, G. (2002). Advantages, opportunities and limits of empirical evaluations: Evaluating adaptive systems. *Künstliche Intelligenz*, *16*(3), 17–20.

307. Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

308. Winter, S. (2001). Ontology: buzzword or paradigm shift in GI science? *International Journal of Geographical Information Science*, *15*(7), 587–590.

309. Wittgenstein, L. (2009 (1953)). *Philosophical Investigations*, 4th edn. Chichester, UK: Blackwell, trans. G. E. M. Anscombe.

310. Woods, W. (1975). What's in a link: Foundations for semantic networks. In: Bobrow, D., Collins, A. (Eds) *Representation and Understanding: Studies in Cognitive Science*, New York: Academic Press, pp. 35–82.

311. Wright, K. (2005). Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication*, *10*(3), URL `http://jcmc.indiana.edu/vol10/issue3/wright.html`, article 11.

312. Wu, D., Zhao, D., Zhang, X. (2008). An Adaptive User Profile Based on Memory Model. In: *Proceedings of the 9th International Conference on Web-Age Information Management, WAIM'08* (pp. 461–468), IEEE.

313. Wu, Z., Palmer, M. (1994). Verbs semantics and lexical selection. In: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, ACL-94* (pp. 133–138), ACL.

314. Wubben, S., van den Bosch, A. (2009). A semantic relatedness metric based on free link structure. In: *Proceedings of the Eighth International Conference on Computational Semantics* (pp. 355–358), ACL.

315. Yeh, E., Ramage, D., Manning, C., Agirre, E., Soroa, A. (2009). WikiWalk: random walks on Wikipedia for semantic relatedness. In: *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing* (pp. 41–49), ACL.

316. Yu, S., Spaccapietra, S., Cullot, N., Aufaure, M. (2004). User Profiles in Location-Based Services: Make Humans More Nomadic and Personalized. In: Hamza, M. (ed.) *Proceedings of the IASTED International Conference on Databases and Applications, DBA'04* (pp. 25–30), Calgary, Canada: ACTA Press.

317. Yu, W., Lin, X., Le, J. (2010). Taming Computational Complexity: Efficient and Parallel SimRank Optimizations on Undirected Graphs. In: *Proceedings of the 11th International Conference on Web-age Information Management, WAIM 2010* (pp. 280–296), Springer, LNCS, vol. 6184.

318. Zhao, P., Han, J., Sun, Y. (2009). P-Rank: A Comprehensive Structural Similarity Measure over Information Networks. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09* (pp. 553–562), ACM.

319. Zimmermann, A., Specht, M., Lorenz, A. (2005). Personalization and context management. *User Modeling and User-Adapted Interaction*, *15*(3), 275–302.

APPENDIX

A

SURVEYS

A.1 Open Source Software Survey

This is the full text of the online questionnaire utilised in our survey of open source Web mapping software (see Section 4.3). This questionnaire was disseminated through online forums and mailing lists in April 2010.

—

We are conducting a survey on Open-Source technologies with a particular focus on Geo-Spatial projects. Our goal is to collect first-hand knowledge about a number of Open-Source projects active on the Internet. With this work we hope to identify strong and weak points of each project in order to give some guidelines for future directions to the Open-Source community and potential developers in relation to Geo-Spatial research.

Therefore we would like to ask you to take an anonymous questionnaire on {**Project name**} that we have included in our survey.

Title: [Survey on Geo-Spatial Open-Source Technologies] Questionnaire on {**Project name**}

1. **Contribution to the project** - Did you actively contribute to the project as a developer?
   No answer yes no

2. **Affiliation with project owners** - Are you a project owner or are you affiliated with the project owners?
   No answer yes no

3. **Level of expertise** - What is your personal experience level with this software?
   No answer 0 Beginner 1 2 3 4 5 Expert

4. **Learning curve** - Do you think it is easy or hard to learn how to use this software as a developer?
   No answer 0 Hard 1 2 3 4 5 Easy

5. **Stability** - Do you think this software is stable and reliable for production use (e.g. few critical bugs, etc.)?
   No answer 0 Unstable 1 2 3 4 5 Stable

6. **Performance** - In your experience, how would you define the performance of the software?
   No answer 0 Slow 1 2 3 4 5 Fast

7. **Scalability** - In your experience, how would you define the scalability of the software?
   No answer 0 Poor 1 2 3 4 5 Excellent

8. **Interoperability** - How would you define the integration of the software with other technologies?
   No answer 0 Hard 1 2 3 4 5 Easy

9. **Extendibility** - Is it easy to extend the software functionalities with external plugins/addons?
   No answer 0 Hard 1 2 3 4 5 Easy

10. **Standards** - How would you define the software support for widely-adopted standards?
    No answer 0 Poor 1 2 3 4 5 Excellent

11. **Documentation** - What do you think of the documentation of the software (e.g. readability, completeness, quality, useful examples, etc.)?
    No answer 0 Poor 1 2 3 4 5 Excellent

12. **Community support** - How would you define the technical support offered on the project forums/mailing lists?
    No answer 0 Poor 1 2 3 4 5 Excellent

13. **Frequency of updates** - How would you define the new releases containing new features, improvements and bug fixes?
    No answer 0 Rare 1 2 3 4 5 Frequent

14. **Comments** - If you have any comments about this survey and/or {**Project name**}, you can put them here.

A.2 Geo Relatedness and Similarity Dataset Online Survey

These two questionnaires were used to collect human judgements respectively on semantic relatedness and similarity for OpenStreetMap (OSM) geographic concepts. The Geo Relatedness and Similarity Dataset (GeReSiD) is based on these judgements (see Section 6.2). The questionnaires were disseminated on forums and mailing lists related to geography, OSM, and Geographic Information Systems (GIS) in February 2012. The first part of the questionnaire, i.e. the introduction and the demographic questions, is identical for both relatedness and similarity.

—

Survey on comparison of geographic concepts

Spatial Information Systems Group in University College Dublin

Dear reader, we are studying how people compare geographic entities that are commonly found in maps. We would greatly appreciate it if you could spend 5 minutes taking this test. Your completion of this task is voluntary. Your responses are anonymous. Please do not write your name anywhere on this form.

1. Please select your AGE GROUP. Choose one of the following answers: $18-25, 26-35, 36-45, 46-55, 56-65, 65+$
2. Is ENGLISH your native language? Yes - No
3. Please indicate your GENDER: Male - Female
4. Please indicate your level of familiarity with WEB MAPS. Choose one of the following answers: Never used - Occasional user - Frequent user - Expert
5. Please indicate your CONTINENT of origin. Choose one of the following answers: Africa - Asia - Europe - North America - South America - Oceania
6. If you have any comments, suggestions or criticism, please enter them here: _____

Relatedness questionnaire

7. This is a list of pairs of geographic concepts. The concepts should be considered in the context of web maps. To what degree are they CONCEPTUALLY RELATED?

   For example, conceptually related concepts are

   APPLES - BANANAS

DOCTOR - HOSPITAL

TREE - SHADE

Choose a value between 1 (very unrelated) and 5 (very related). If the meaning of a concept in a pair is unclear, select 'no answer'. Please bear in mind that this is not a test that has right or wrong answers.

(*Randomised list of 50 pairs of geographic concepts*)

basketball court $\leftrightarrow$ volleyball facility: $1-2-3-4-5-$ no answer
industrial landuse $\leftrightarrow$ landfill: $1-2-3-4-5-$ no answer
theatre $\leftrightarrow$ cinema: $1-2-3-4-5-$ no answer
...
canal $\leftrightarrow$ dock: $1-2-3-4-5-$ no answer

Similarity questionnaire

7. This is a list of pairs of geographic concepts. The concepts should be considered in the context of web maps. To what degree are they CONCEPTUALLY SIMILAR?

For example, conceptually similar concepts are

APPLES - BANANAS

DOCTOR - SURGEON

CAR - MOTORCYCLE

Choose a value between 1 (very dissimilar) and 5 (very similar). If the meaning of a concept in a pair is unclear, select 'no answer'. Please bear in mind that this is not a test that has right or wrong answers.

(*Randomised list of 50 pairs of geographic concepts*)

basketball court $\leftrightarrow$ volleyball facility: $1-2-3-4-5-$ no answer
industrial landuse $\leftrightarrow$ landfill: $1-2-3-4-5-$ no answer
theatre $\leftrightarrow$ cinema: $1-2-3-4-5-$ no answer
...
canal $\leftrightarrow$ dock: $1-2-3-4-5-$ no answer

APPENDIX

B

WORDNET-BASED MEASURES

This Appendix details the workings of ten WordNet-based similarity/relatedness measures, summarised in Section 2.5.4.

[**Edge count (path)**] This is a simple measure of similarity based on edge counting. Its application to semantic networks is discussed by Rada et al. [239].

$$sim_{path}(t_a, t_b) = \frac{1}{len(t_a, t_b)} \tag{B.1}$$

where len is the length of the shortest path, considering is_a relationships. This measure works only for nouns and verbs.

[**Leacock and Chodorow (lch)**] The measure devised by Leacock and Chodorow [170] finds the path length between t_a and t_b in the is_a hierarchy in WordNet. The path length is scaled by the maximum depth of the hierarchy in which they reside (D).

$$sim_{lch}(t_a, t_b) = -log\frac{len(t_a, t_b)}{2\ D} \tag{B.2}$$

The measure is very similar to *path*.

[**Resnik (res)**] Resnik [245] developed the first relatedness measure that combines knowledge-based and corpus-based approaches. According to the author, the "more information two concepts share in common, the more similar they are, and the information shared by two concepts is indicated by the information content of the concepts that subsume them in the taxonomy" (p. 449). Therefore, his measure of similarity between t_a and t_b is the information content (I_c) of the most specific term that both terms have in common, i.e. their lowest common subsumer ($lcs(t_a, t_b)$) in the is_a WordNet hierarchy:

$$sim_{res}(t_a, t_b) = max\ \{I_c(lcs(t_a, t_b))\} \qquad I_c(t) = -log\ p(t, C) \tag{B.3}$$

where $p(t, C)$ is the probability to encounter term t in corpus C. When the terms have more than one lcs, that with maximum I_c is chosen.

[**Jiang and Conrath (jcn)**] Jiang and Conrath [137] have extended Resnik's measure, taking into account the information contents of t_a and t_b

as well as their lowest common subsumer ($lcs(t_a, t_b)$). They define a distance function as follows:

$$dist_{jcn}(t_a, t_b) = I_c(t_a) + I_c(t_b) - 2\ max\ \{I_c(lcs(t_a, t_b))\} \qquad (B.4)$$
$$sim_{jcn}(t_a, t_b) = 2\ D - dist_{jcn}(t_a, t_b)$$

where D is the maximum possible path length.

[Lin (lin)] Lin [180] proposed a universal definition of similarity in terms of information theory. The fundamental similarity theorem states that the "similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B." Formally:

$$sim(A, B) = \frac{log\ p(common(A, B))}{log\ p(description(A, B))} \qquad (B.5)$$

When applied to semantic networks, this similarity theory is strongly associated to those of Jiang and Conrath, and Resnik:

$$sim_{lin}(t_a, t_b) = \frac{2\ max\ \{I_c(lcs(t_a, t_b))\}}{I_c(t_a) + I_c(t_b)} \qquad (B.6)$$

[Wu and Palmer (wup)] As part of their study on machine translation of verbs, Wu and Palmer [313] proposed an approach to semantic similarity based on the edge count between the lowest common subsumer (lcs) and the terms:

$$lcs_{ab} = lcs(t_a, t_b) \qquad len_{root} = len(lcs_{ab}, t_{root}) \qquad (B.7)$$
$$sim_{wup}(t_a, t_b) = \frac{2\ len_{root}}{len(t_a, lcs_{ab}) + len(t_b, lcs_{ab}) + 2\ len_{root}}$$

where t_{root} is the root node of the hierarchy.

[Hirst and St-Onge (hso)] Hirst and St-Onge [124] described a 'lexical chain' as "a cohesive chain in which the criterion for inclusion of a word is that it bears some kind of cohesive relationship (not necessarily one specific relationship) to a word that is already in the chain" (p. 307). In practice, lexical chains are directed graphs that are constructed and updated on the basis on each term's relationships in WordNet. As new terms are inserted into it, a lexical chain becomes semantically narrower and less ambiguous, strengthening relevant edges and removing irrelevant ones.

This measure classifies relationships in WordNet as having a direction, i.e. generalisations are going up, while specialisations are going down. It then establishes the relatedness between two terms by finding a path through a trade-off between changes of direction and length. So for all the existing paths between t_a and t_b (P), the measure is computed through a weight function w:

$$
\begin{aligned}
P &= \{p_0, p_1, \ldots p_n\} \\
w(p_{ab}) &= A \cdot len(t_a, t_b) \cdot B \cdot \delta(t_a, t_b) \\
sim_{hso}(t_a, t_b) &= max\{w(p_{ab}) : p_{ab} \in P)\}
\end{aligned}
\tag{B.8}
$$

where $\delta(t_a, t_b)$ is the number of changes of direction to reach t_b from t_a. A and B are constants.

[Banerjee and Pedersen (lesk)] Banerjee and Pedersen [28] have adapted the classic Lesk algorithm to WordNet. The Lesk [176] algorithm was originally developed to disambiguate word senses in short sentences, i.e. as the author put it, "how to tell a pine cone from an ice cream cone" (p. 24). To do so, the algorithm starts by extracting the definitions ('gloss') of all the words in the sentence, for each sense of each word. Subsequently, it computes the set of overlapping words between a target word and all the other words in the sentence. The sense having the largest overlap is selected as the most suitable. The authors call this measure 'extended gloss overlap.' The relatedness of two terms is therefore proportional to the overlap between their glosses:

$$
sim_{lesk}(t_a, t_b) = gloss(t_a) \cap gloss(t_b) \tag{B.9}
$$

The set $gloss(t)$ is constructed by visiting the related terms linked to t through semantic relationships in WordNet.

[Patwardhan and Pedersen's gloss vector (vector)] Schütze [260] devised 'context vectors' (second order co-occurrence vectors) to perform word sense disambiguation. Context vectors were adopted for WordNet by Patwardhan and Pedersen [232], representing the word senses by second order co-occurrence vectors of their glosses, using WordNet glosses as a text corpus. The relatedness of the two senses is computed as the cosine distance of their gloss vectors:

$$
sim_{vector}(t_a, t_b) = 1 - cos(\vec{c_v'}(t_a), \vec{c_v'}(t_b)) \tag{B.10}
$$

Second order co-occurrence vector $\vec{c_v'}(t)$ is built by collecting terms contained in t gloss, and extending it to terms linked to t. First order co-occurrences are terms that occur near each other in the corpus (used to build vector $\vec{c_v}$), while second order co-occurrences are terms that are first order co-occurrences to the same term.

[Patwardhan and Pedersen's paired gloss vector (vectorp)] This is an extension of the *vector* measure discussed by Patwardhan and Pedersen [232], which relies on second order co-occurrence vectors to compute semantic relatedness. In this approach, the gloss vectors of terms t_a and t_b are expanded by including glosses from related terms. Given a set of semantic relationships R, a pair of context vectors is generated for each relationship $r \in R$. The overall relatedness is the sum of all the cosine distances:

$$R = \{r_0, r_1 \dots r_n\} \tag{B.11}$$

$$sim_{vectorp}(t_a, t_b) = \sum_{i=0}^{|R|} 1 - cos(\vec{c_{ri}}(t_a), \vec{c_{ri}}(t_b))$$

where $\vec{c_{ri}}(t)$ is a second order co-occurrence vector constructed on t and all the terms related to t through semantic relationship r.