



Geographic Information Retrieval

Danilo Montesi
Yisleidy Linares Zaila



Outline

1. Introduction
 - Geographic Information Retrieval (GIR): basic concepts
 - General architecture
2. Geographic Indexes
 - Separate Index
 - Hybrid Index
 - Use of R-tree structure
3. GIR resources
 - Gazetteers
 - Geographic ontologies
4. Geographic Information Extraction
 - Toponym Recognition
 - Toponym Disambiguation
5. Geographic Data Representation
 - Geographic coordinates
 - Minimum Bounding Rectangle



Outline

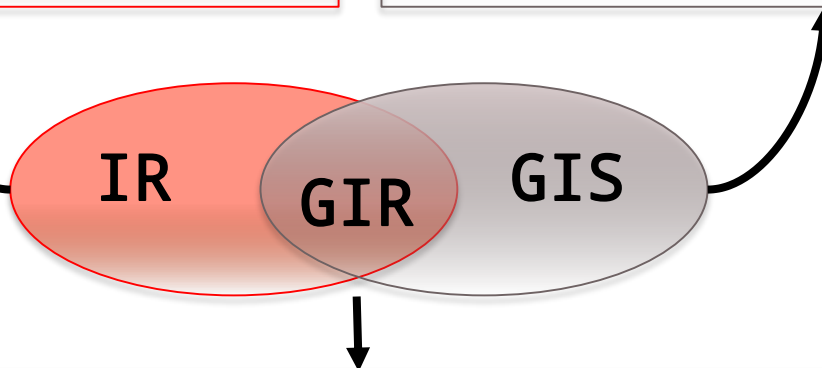
6. Query Processing
 - Query Expansion
7. Geographic Similarity Measures
 - Euclidean distance
 - Topological relationship
8. Textual and Geographic Rankings
 - First textual, after geographic or vice-versa
 - Linear combination
9. GIR evaluation
 - GeoCLEF test collections
 - Evaluation Measures
10. IR frameworks
 - Terrier
11. Current challenges



Geographic Information Retrieval

Finding information resources (**usually documents**) of an unstructured nature (**usually text**) relevant to an information need (**user query**) from a large collection (**e.g. WEB**) [1]

Development and uses of **theories, methods, technology** and **data** for understanding **geographic processes, relationships** and **patterns** [2]

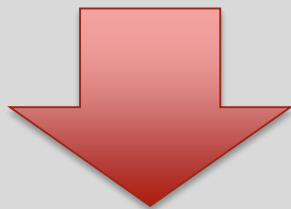


Techniques to build an application system that could well **index, query, retrieve** and **browse** the **geo-referenced information**. GIR is supposed to be able to **better understand** the **geographical knowledge** contained within **web documents** and **user queries**, and provide a more satisfactory answer to user needs.

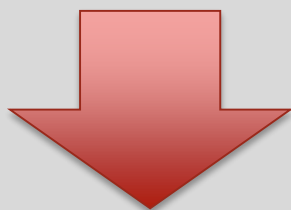


GIR Concepts

Document Collection



Indexes



Results



Information Extraction (IE)

- Extracting information from unstructured documents
- Representing it with appropriate descriptors
- Organizing it into structured indexes

Supported by **NLP tasks**

Information Retrieval (IR)

- Query **information extraction**
- Matching user needs with indexed document descriptors



Natural Language Processing (NLP)

NLP is a field of computer science, artificial intelligence, and computational linguistics focused on developing efficient algorithms to process unstructured texts and to make their information accessible to computer applications.

Most common NLP tasks:

- Tokenization
- Part-of-Speech tagging
- Name Entity Recognition

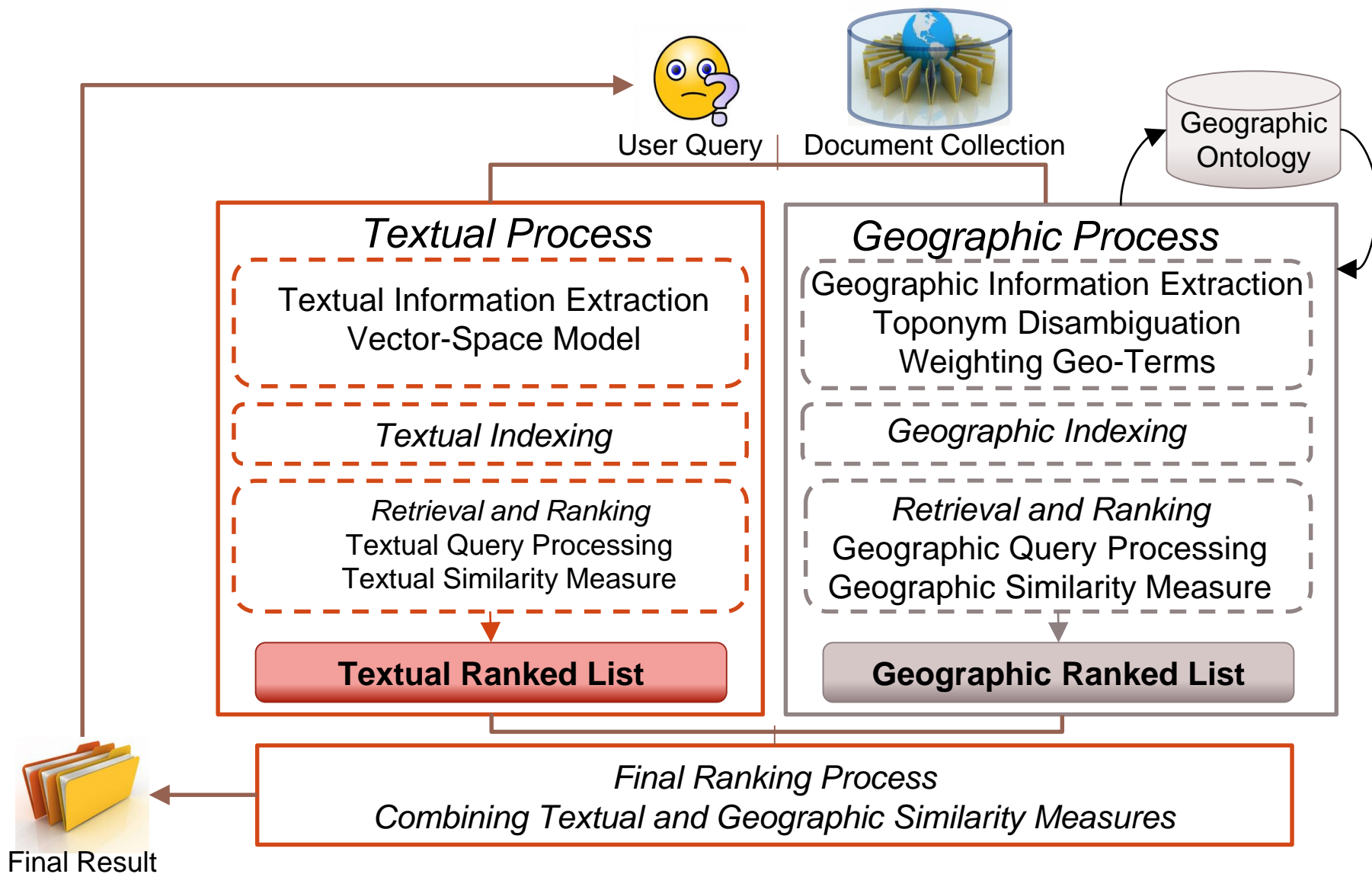
LingPipe¹ and OpenNLP² are NLP platforms that support **spatial named entity recognition** from textual documents.

¹ <http://alias-i.com/lingpipe/>

² <https://opennlp.apache.org/index.html>

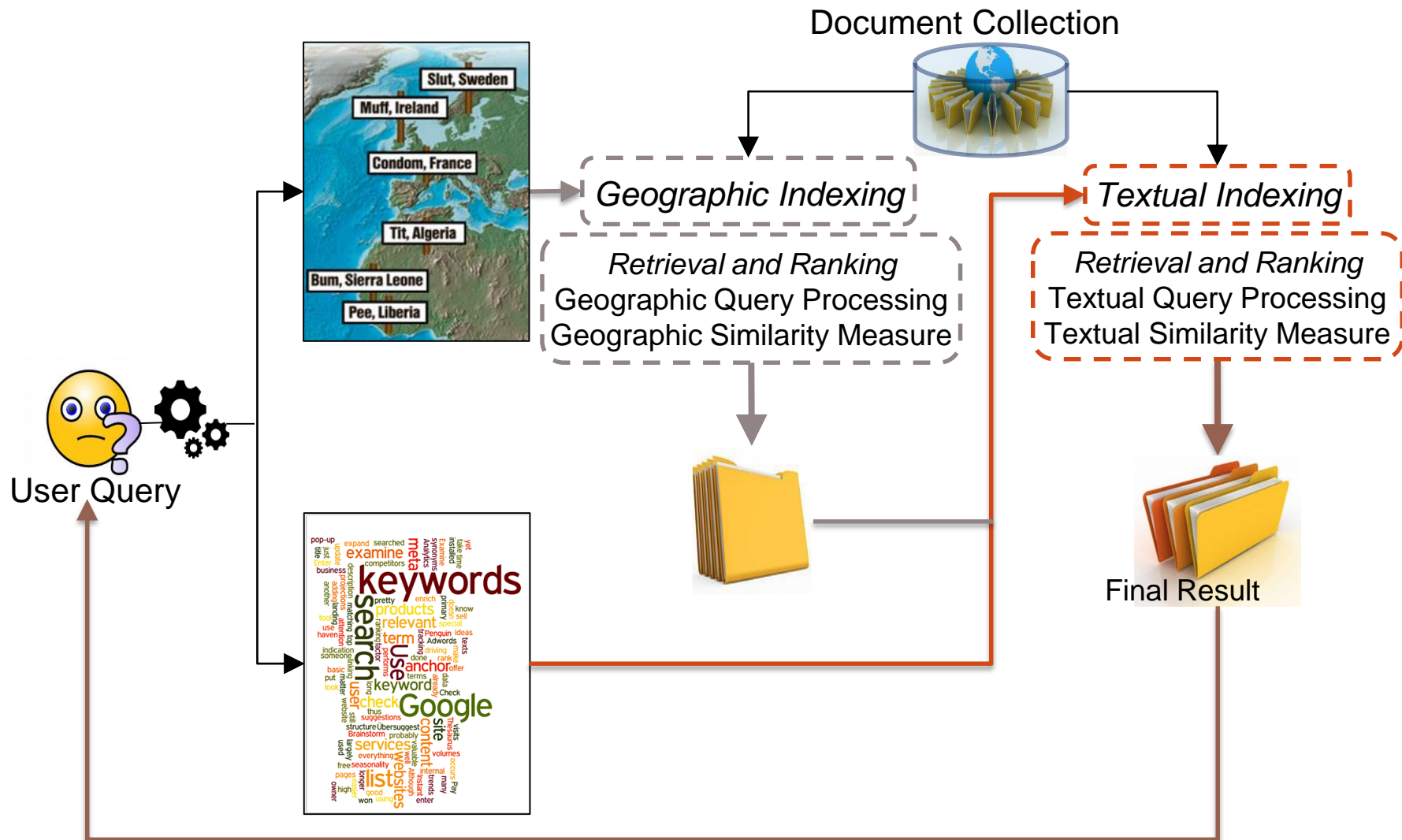


General GIR architecture



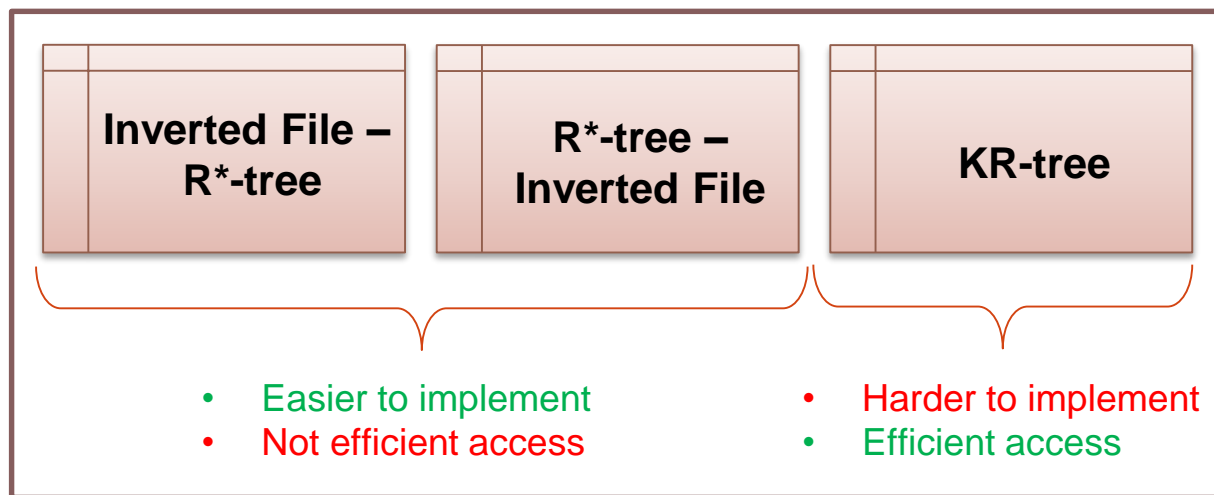
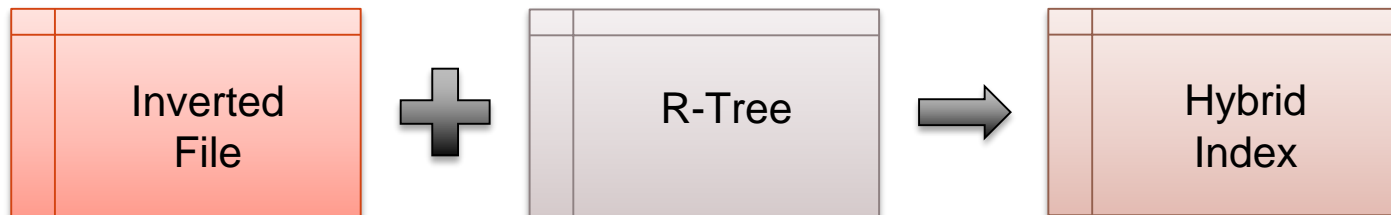


Separate Index (2/2)

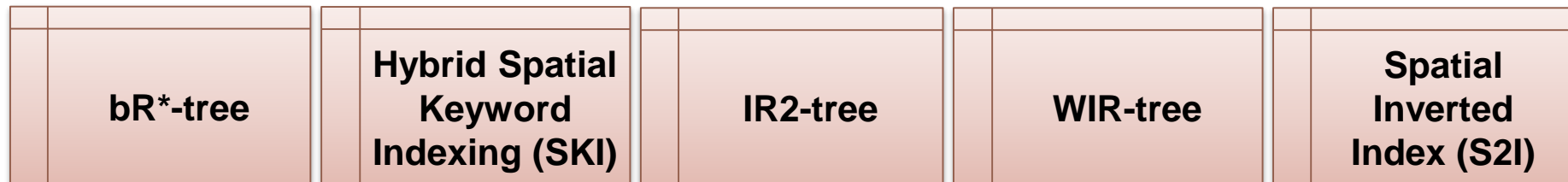




Hybrid Index

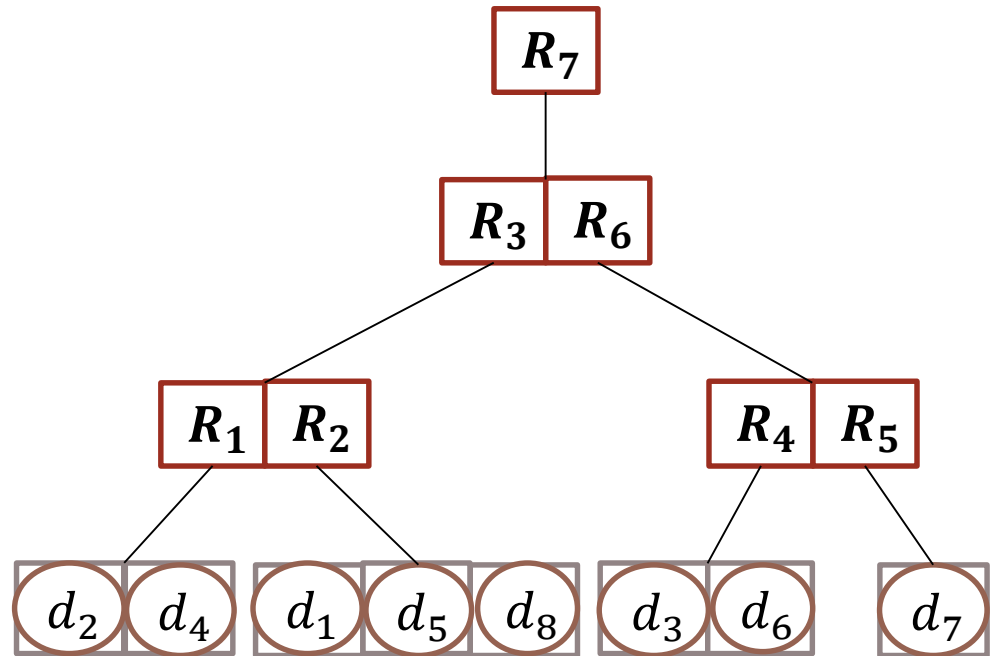
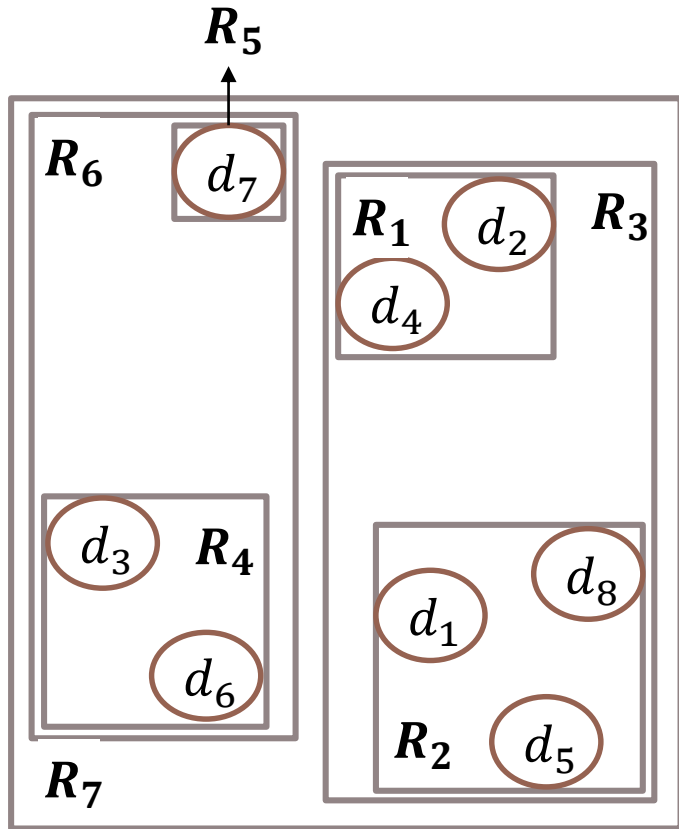


Other indexes...



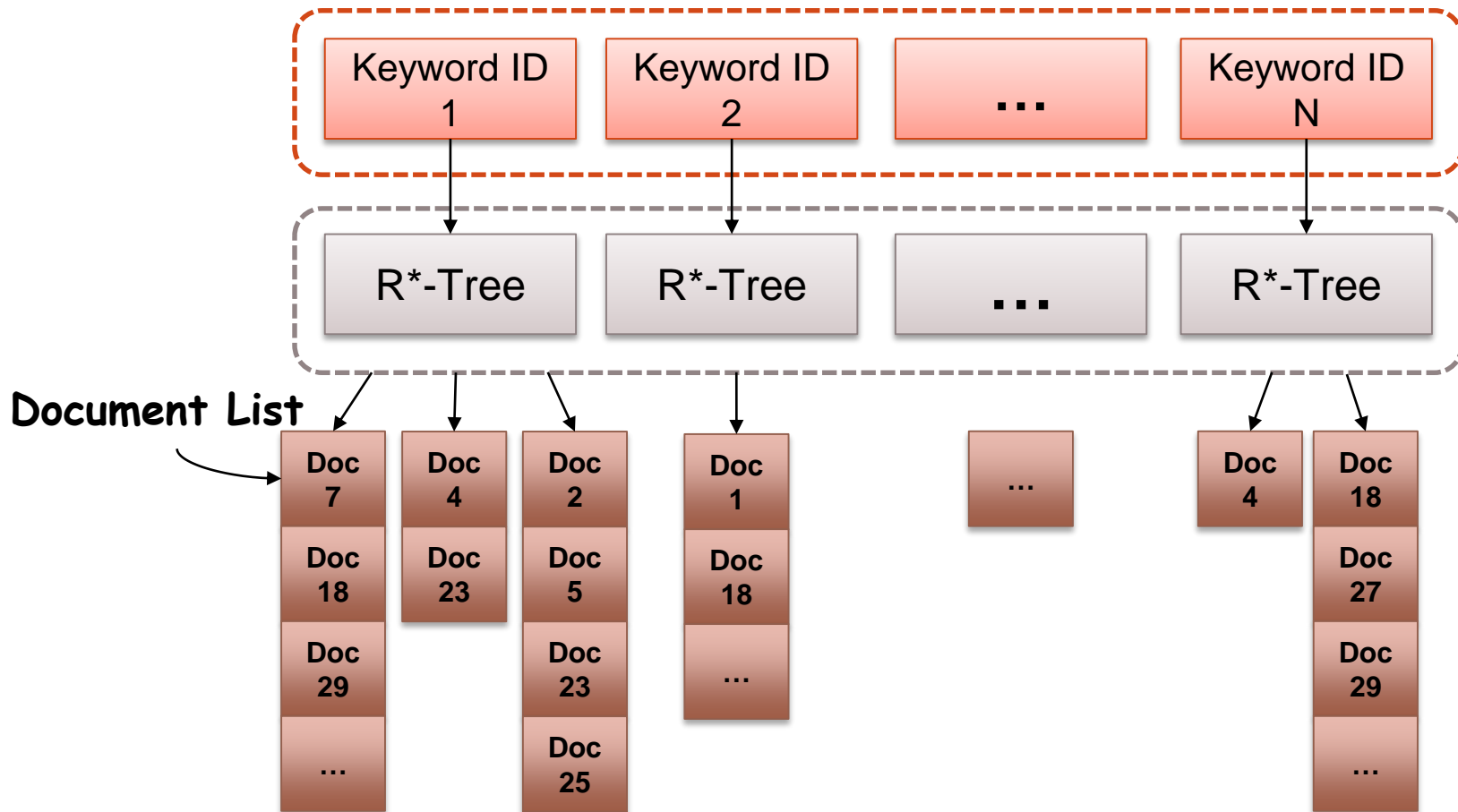


R*-Tree data structure representation

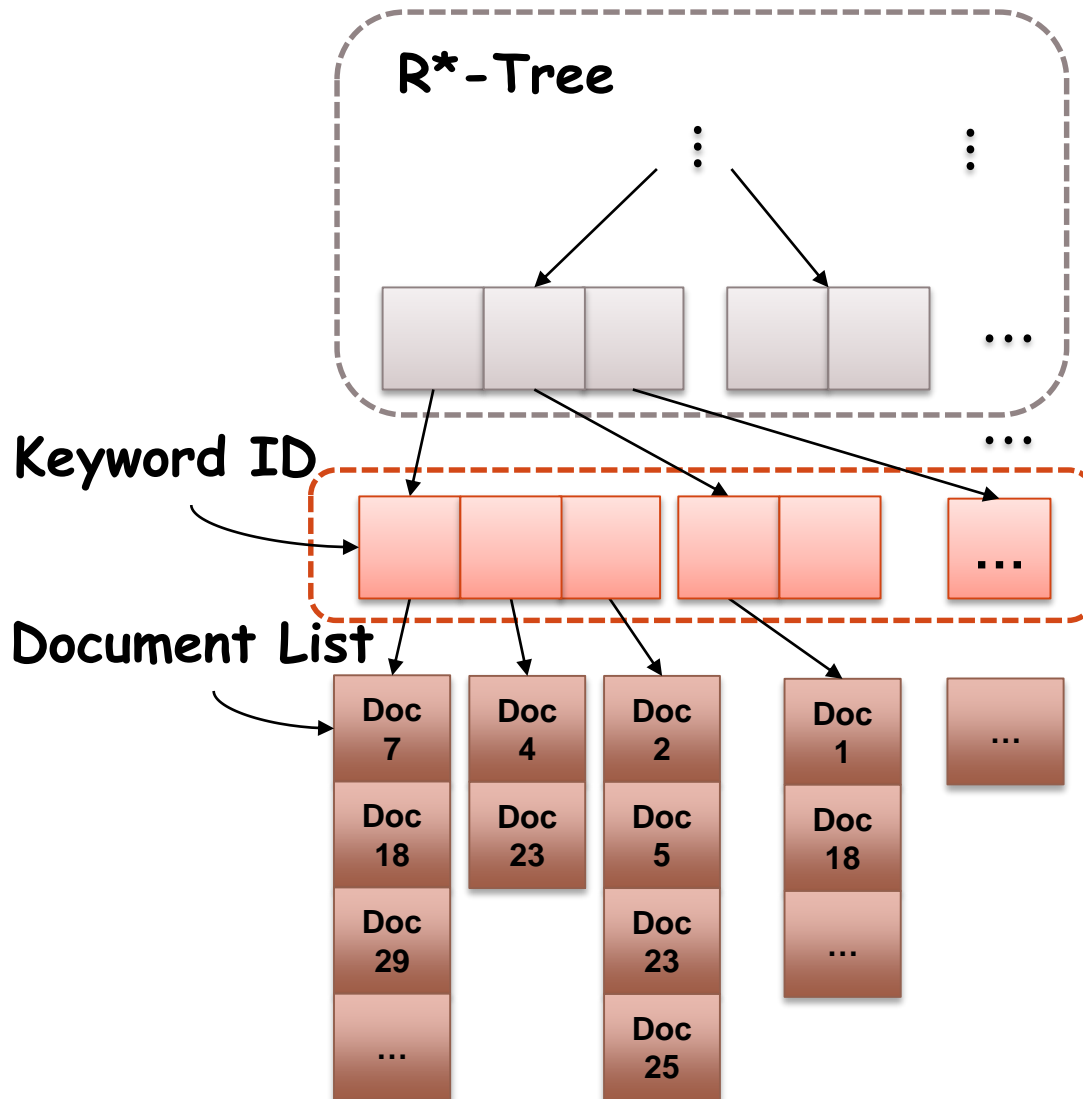




Inverted File – R*-Tree



R*-Tree – Inverted File





KR*-tree (*Keyword-R*-tree*) (1/2)

Inverted-File -- R*-tree & R*-tree – Inverted File

Keywords are maintained
separately



Queries are answered by the
intersection of documents
IDs from the inverted file or
R*-trees

KR*-Tree

Capture the joint
distribution of keywords



Documents IDs containing
the query keywords are
directly obtaining **without**
merging any lists

Greatly enhance the performance



KR*-tree (*Keyword-R*-tree*) (2/2)

KR*-tree is built in a way similar to an R*-Tree, but with ***minimal overhead in handling the keywords***

Two distinct steps

1. The set of keywords corresponding to each node of the tree is determined
2. The set of keywords in each node is converted to a KR-Tree list

KR-tree list is similar to that of an inverted file index, but ***it stores the node instead of document IDs***

The number of internal nodes of KR*- tree is considerably smaller compared to the number of documents indexed



Resources: Gazetteers

A **gazetteer** is an alphabetical list of place names with information that can be used to locate the areas that the names are associated with

Alphabetical List

Includes place names and locations as well as information typically found in atlases

Dictionary

Includes location information in the form of geographic coordinates or descriptions of spatial relationships to other places

Encyclopedic

Includes all the information of a dictionary but the information is more detailed and may come in the form of articles

Getty Thesaurus



Canadian Geographical Names

Natural Resources
Canada



Place Name Sites



GEOnet Names



Gazetteers examples



Resources: Geo-Ontologies (1/4)

Ontology

Theory that uses a specific vocabulary to describe entities, classes, properties, and function related to a main view of the world [20]

Geographic Ontology

*An ontology with **spatial relationships** among **geographic features**. A geo-ontology describes:*

- Entities that can be assigned to locations on the surface of the Earth*
- Semantic relations among these entities that include, hypernymy, hyponymy, mereonymy and synonymy relation*
- Spatial relations among entities (e.g. adjacency, spatial containment, proximity, connectedness)*

Ontologies are vital to semantic description and referencing of geographic information. They contain in a very structured way, domain knowledge and specific data regarding a certain subject field [19]



Resources: Geo-Ontologies (2/4)

Basic Design Idea of Ontology-based IR

1. Establish the ontology of the related fields

2. Collect data from the information sources and to store it in **prescribed format**

3. In accordance with the ontology, a query converter transforms the search request from user interface into **prescribed format**

4. After processing, the result of the retrieval is returned to the user

Resources: Geo-Ontologies (3/4)



Yahoo! GeoPlanet

- . Uses web services to unambiguously geotag data across the web
- . + 6 million named places globally
- . uses a hierarchical model which preserves geographical relationships



GeoWordNet

- . WordNet + GeoNames + Italian part of MultiWordNet
- . + 3.5 million of entities.

Resources: Geo-Ontologies (4/4)



GeoNames

- . Crowd sourced open project.
- . +10 million geographical names and +8 million unique features.
- . Geographical data includes names of places in various languages, elevations, population and others, from various sources.
- . Users may manually edit, correct and add new names using a user friendly wiki interface.



Geographic Information Extraction (1/6)

Identification of geographic terms in documents and queries, and associating these terms with the appropriate geographic locations

Geographical terms participate in the definition of the geographical scopes of documents and queries.

There are four basic concepts relevant for GIR systems:

1. **Place-names** (e.g. Italy, Miami, Toronto, etc.)
2. **Geographical relations** (e.g. south, north, near, far, etc.)
3. **Geographical concepts** (e.g. lakes, cities, mountains, etc.)
4. **Geographical adjectives** (e.g. Italian, Canadian, northern, etc.).



Geographic Information Extraction (2/6)

Assigning geographic scopes to documents include:

- ***correct identification of the geographic terms***
- ***toponym disambiguation problem*** (e.g. *Havana could be the capital of Cuba or part of the name of famous Cuban rum Havana Club; or London could be the capital of England or a city in Ontario*)

Toponym disambiguation is currently a top challenge in GIR

Leidner [4]

Woodruff and Plaunt [29]

Ding [12]

Martins [31]

Amitay [30]

Campelo and Baptista [29]



Geographic Information Extraction (3/6)

Woodruff and Plaunt [5]

- Compute the geographic scope of a document based on the references that appear in the text.
- The method is based on disambiguating the geographic references into their respective bounding polygons.
- The geographic scope of the document is computed using the overlapping area for all the polygons, trying to find the most specific place that is related to all the place references mentioned in the text.

Ding [6]

- It is also a technique based on the geographical scope.
- The geographical scope of a web resource ω is defined with the help of a hierarchical gazetteer.
- It introduces two concepts: Power and Spread to decide the geographical scope the creator of web page intends to reach.
- They proposed two kinds of resources for the calculation of a web page's geographical scope: the locations of the pages linking to web resource ω and the placenames appearing in ω



Geographic Information Extraction (4/6)

Martins [7]

- It uses the PageRank algorithm to infer a single global scope for each document.
- First, geographical references are extracted from the text and associated with the corresponding concepts in a geographical ontology.
- Every document is represented by a set of features corresponding to geographical references from the text.
- Each feature associates a weight to a set of concepts in the ontology according to their occurrence frequency.
- A graph is used to represent the association between the geographical references and the ontology concepts in order to apply the PageRank algorithm to infer geographic scope.
- Each concept is represented as a node in the graph, and a relationship statement is represented by two directed edges.
- Different types of relationship in the ontology correspond to different edge weights in the graph.
- A PageRank formula based on edge and node weights is used to compute the ranking score for each node in the graph. The placename with the highest weight is finally assigned as the geographical scope of the page.



Geographic Information Extraction (5/6)

Amitay [8]

- Amitay proposed a system named Web-a-Where.
- It uses a hierarchical gazetteer as their knowledge base.
- Firstly they generate a hierarchical relationship for every placename appearing in the document as A/B/C (e.g. New York/USA/North America).
- Assigning to each node its value of importance, they sort the nodes by these values and then select the most relevant placenames as the geographical focus of the document.

Leidner [9]

- A document's scope is defined at country level.
- They employed the country of the document's publication and the countries of the most important unambiguous placenames extracted from the text.



Geographic Information Extraction (6/6)

Campelo and Baptista [10]

- They proposed a model for detecting geographic references in Web Documents based on a set of heuristics.
- They introduced the concepts of confidence factor and confidence modifier, which aim to measure the probability of a detected geographic reference of being a valid reference and being associated to a correct place, even when there exists ambiguity.

Query Processing (1/2)

*In a GIR system, the analysis of a query should answer the following questions: **what to search?**, **where to search?** and **what is the relationship between what and where?***



<what, relation, where>



<what, relation, where>
<what, relation, where>
< ... , ... , ... >
<what, relation, where>

query expansion aims to make the user query resemble more closely the documents it is expected to retrieve.

Query expansion refers to content and typically is limited to adding, deleting or reweighting of terms.



Query Processing (2/2)

Cardoso [11] presents an approach for geographical query expansion based on the use of feature types, readjusting the expansion strategy according to the semantics of the query.

Buscaldi [12] uses WordNet during the indexing phase by adding the synonyms and the holonyms of the encountered geographical entities to each documents index terms, proving that such method is effective.

Stokes [13] concludes that significant gains in GIR will only be made if all query concepts (not just geospatial ones) are expanded.

These results show that the development of algorithms based on query expansion concepts seems to be a good start point for tackling query processing phase in GIR systems.



Geographic Data Representation (1/2)

Geographical data can be represented as:

- ***Place names***: Qualitative descriptors of the geographic extent
- ***Spatial reference system***: Projection and coordinate system information
- ***Spatial Representation Model***: Vector, raster
- ***Spatial features***: Type and quantity
- ***Minimum Bounding Rectangle (MBR)***: A coordinate pair defining a MBR



Geographic Data Representation (2/2)

MBRs

Advantages

- Provide a compressed, abstract approximation of spatial object.
- The representation is conceptually powerful because it evokes a printed map.
- Its simplicity, computational efficiency and storage advantages make it the most commonly used spatial approximation. [14]

Disadvantages

- Weaknesses when representing diagonal, irregular, non-convex or multi-part regions [15]
- Over-estimate area, misrepresenting shape, and fail to capture the distribution of the data within themselves, leading to *false positives* in GIR matching

MBR still represents state of the art for GIR applications



Geographic Similarity Measures

Larson and Frontiera [16], as well as Andogah [17] review **major techniques for measuring similarities**.

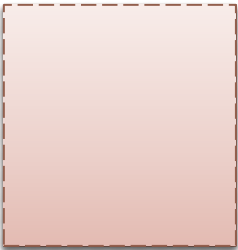
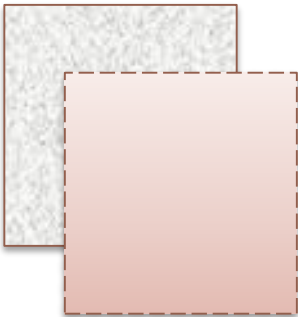
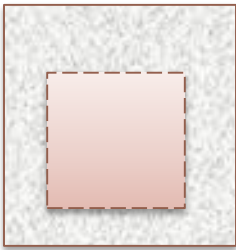

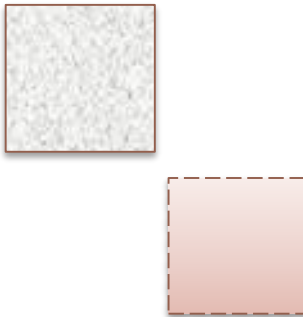
- ***Euclidean Distance*** measures the proximity between document scope and query scope
- ***Extent of overlap*** measures the proportion of overlap between document scope and query scope; the greater overlap, the higher the relevance of the document
- ***Containment relation*** ranks documents by ratio of document scope to query scope (e.g., when the document scope is inside the query scope) or by ratio of query scope to document scope (e.g., when the query scope is inside the document scope)



Geographic Similarity Measures

Spatial similarity based on overlapping regions	
Reference	Formula
Hill [18]	$SimG = 2 * O / (Q + C)$
Walker [19]	$SimG = \min(O / Q, O / C)$
Beard and Sharma [20]	<p>Case 1: <i>Q contains C</i></p> $SimG = C / Q$ <p>Case 2: <i>Q and C overlap</i></p> $SimG = \frac{O / Q \%}{(1 - O / C \% + 100)}$ <p>Case 3: <i>Q contained in C</i></p> $SimG = Q / C$
<p>Where:</p> <p><i>Q = area of query region</i></p> <p><i>C = area of candidate document region</i></p> <p><i>O = area of overlap for C, Q</i></p> <p>SimG (for all):</p> <p><i>0 = no similarity</i></p> <p><i>1 = identical</i></p>	

Geographic Similarity Measures

Spatial Relationships between overlapping regions				
Equals	Overlaps	Contains	Contained by	Disjoint
				

For geospatial searches, the query is compared with MBRs of all candidate documents using polygon-polygon geometric operations. If there is overlap between the query and the document regions, the document is considered a match.



Geographic Similarity Measures

- GIR ranking methods are based on quantifying the similarity between the query and a document in the collection.
- This similarity score can be interpreted as an estimate of the relevance of a candidate document for a user's information need.
- Retrieved documents are ranked and presented to the user in descending order of these scores.

While traditional IR scores are based on the statistical properties of terms in a collection, GIR relies on spatial scores and rankings are based on geospatial characteristics such as size, shape, location and distance.



Geographic Similarity Measures

- Maron and Kuhns [21] first introduced the idea that, given the imprecise and incomplete ways in which a user's information need is represented by a query and an information document by its indexing, **relevance should be approached probabilistically.**
- This is especially true for geographic information retrieval since **all geographic information objects are abstract**, compressed representations of real world phenomena that contain some degree of error and uncertainty.
- In the logistic regression (LR) model of IR, **the estimated probability of relevance for a particular query and a particular document $P(R|Q, D)$ is calculated as the *log odds* of relevance $\log O(R|Q, D)$ and converted from odds to a probability**



Geographic Similarity Measures

The LR model provides estimates for a set of coefficients, c_i , associated with a set of S statistics, X_i , derived from the query and document collection, such that:

$$\log O(R|Q, D) = c_0 \sum_{i=1}^S c_i X_i$$

where c_0 is the intercept term of the regression. The spatial probability of relevance, can be given as:

$$P(R|Q, D) = \frac{e^{X_1 \log O(R|Q, D)}}{1 + e^{\log O(R|Q, D)}}$$

Maroon defined the following geospatial features variables for their logistic regression model:

X_1 = area of overlap(query region, doc region)/ area of query region

X_2 = area of overlap(query region, doc region)/ area of doc region

X_3 = $1 - \text{abs}(\text{fraction of query region that is onshore} - \text{fraction of doc region that is onshore})$



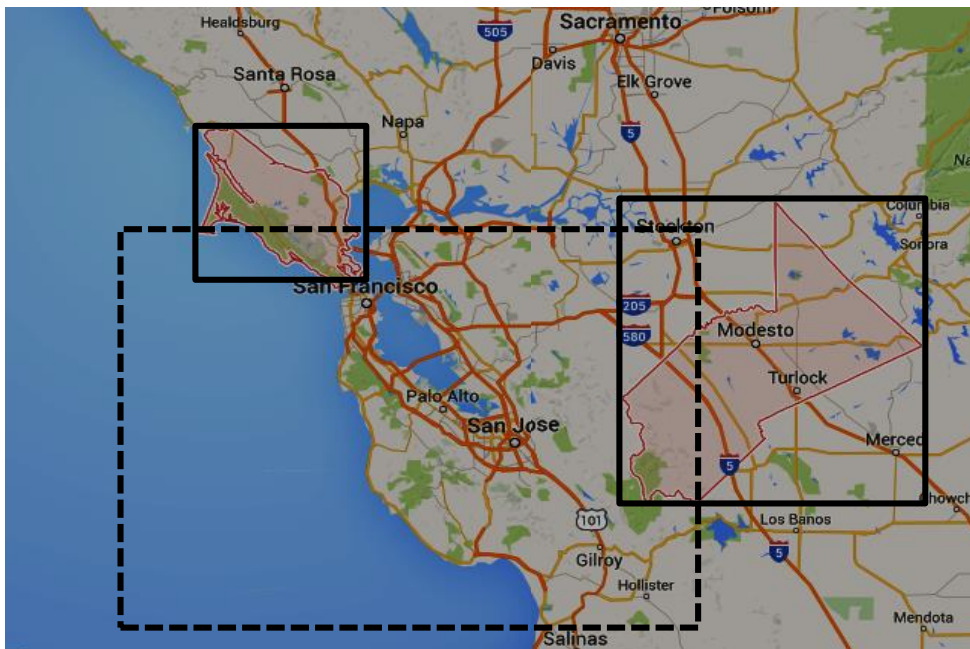
Geographic Similarity Measures

Areas that are near coastline can be problematic when approximated by simplified geometries like the MBR.

The MBR for an offshore region may necessarily include a lot onshore area, and vice versa.

X_3 is defined as a shore-factor variable that captures the similarity between the fraction of a query region that is onshore compared to that of candidate document regions.

Geographic Similarity Measures



- Marin County (geographic scope of document D_1) is **70% onshore**
- Stanislaus County (geographic scope of document D_2) is **100% onshore**.
- Query box (geographic scope of query Q) is **45% onshore**.

Then the shore-factor X_3 for D_1 and D_2 are:

$$X_3(D_1) = 1 - \text{abs}(.45 - .70) = .75$$

$$X_3(D_2) = 1 - \text{abs}(.45 - 1.0) = .45$$



Geographic Similarity Measures

Yong Gao [22] proposed a qualitative approach for supporting geographic information retrieval based on qualitative representation, semantic matching and qualitative reasoning

Information in documents and user queries are represented using propositional logic, which considers the thematic and geographic semantics synthetically.

Thematic information is represented as thematic propositions on the base of domain ontology

Spatial information is represented as geo-spatial propositions with the support of geographic knowledge base.



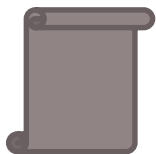
Geographic Similarity Measures

Similarity is based on evidence theory and fuzzy logic.
It is divided into thematic and spatial similarity

Thematic is calculated by the weighted distance of proposition keywords in the domain ontology

Spatial is further divided into conceptual and spatial similarity

Two step process



Document d ,
composed by
 n information
units

1

get $Ru(u_i, q)$ as the combination of **thematic** and **geographic** similarity between u_i and q .

2

get the relevance of document d with q , denoted as $Rq(d, q)$:

$$Rq(d, q) = k(Ru(u_1, q), Ru(u_2, q), \dots, Ru(u_n, q))$$



Geographic Similarity Measures

In this method, two kinds of similarities are combined in different ways:

- A document is a set of information units taken as evidences, so that their similarities are combined by evidence theory.
- The thematic and spatial information of information units, and the query are all represented by proposition logic. Thus, **fuzzy logic reasoning** for the similarity combination function is introduced

Similarity between an information unit and a query: since a query can be represented as a simple information unit, the similarity $Ru(u_i, q)$ between an information unit u_i and a query q can be converted into the similarity between two information units.

Each information unit is composed of thematic and spatial information, so the similarity measurement can be divided into three parts: thematic similarity, spatial similarity and their combination.



Geographic Similarity Measures

- The spatial similarity between two information units is the semantic relevance of their spatial information.
- The spatial information in information units is a sentence of atomic geo-spatial propositions, so the similarity is also a function of them

Suppose p and o are two atomic geo-spatial propositions, and their relevance operator is \odot , then **according to fuzzy logic**, there will be:

$$\begin{cases} p \odot o = g(p, o) \\ p \odot \neg o = 1 - g(p, o) \end{cases}$$

Where $g(p, o)$ is the spatial similarity function.

The spatial similarity function is directional, thus the operator \odot does not satisfy commutative law.



Geographic Similarity Measures

The function $g(p, o)$ is the key problem.

- It can be determined by the belief between atomic geo-spatial propositions, which is inferred based on geographic knowledge base and qualitative spatial reasoning.
- A geographic knowledge base, generally, store the name, geographic extent, and entity type of geographic entities. It also stores spatial relationships explicitly, especially the whole-part relationship.
- Geographic knowledge base is the fundamental of spatial reasoning and computing.
- Geospatial propositions in a document mostly are place names, some of which are constrained by spatial predicated. The same is true for geographic information queries.

The similarity measurement of geographic information turns to the semantic similarity between two place names.



Geographic Similarity Measures

- The semantics of place names includes conceptual and locational features, so the semantic between place names should take both features into account.
- Conceptual similarity is determined by:

$$CS(p, o) = 1 - \left(\sum_{x \in \{p.PartOf - o.PartOf\}} \frac{\alpha}{L_x} + \sum_{y \in \{o.PartOf - p.PartOf\}} \frac{\beta}{L_y} + \sum_{z \in \{p, o\}} \frac{\gamma}{L_z} \right)$$

Where:

- L_x , L_y and L_z are the depths of the sets x, y, z in the ontology respectively.
- The sets of terms $p.PartOf$ and $o.PartOf$ refer to the transitive closure of the parents of p and o respectively in the ontology.
- The weights α, β, γ are harmonic coefficients, providing control over the application of the measure.



Geographic Similarity Measures

- The locational similarity is measured based on the rule “topology matters, metric refines” [23].

Topology similarity

$$Inside(p, o) = \begin{cases} \frac{NumDescendant(o)+1}{NumDescendant(p)+1}, & o \subseteq p \\ 0, & \text{others} \end{cases}$$

Metric similarity

$$Proximity(p, o) = \frac{1}{1+Distance(p,o)/diagonal(p)}$$

Siblings function

$$Siblings(p, o) = \begin{cases} 1, & \text{if } parent\ of\ p = parent\ of\ o \\ 0, & \text{otherwise} \end{cases}$$

- Combine the three values for generating the locational similarity

$$LS(p, o) = b \times \{Inside(p, o) + Proximity(p, o)\} + (1 - b) \times Siblings(p, o)$$

where b is a harmonic coefficient ranging from 0 to 1



Geographic Similarity Measures

- The two similarities, LS and CS, can be combined into a weighted function as the spatial similarity function as:

$$g(p, o) = w_g \times LS(p, o) + w_h \times CS(p, o)$$

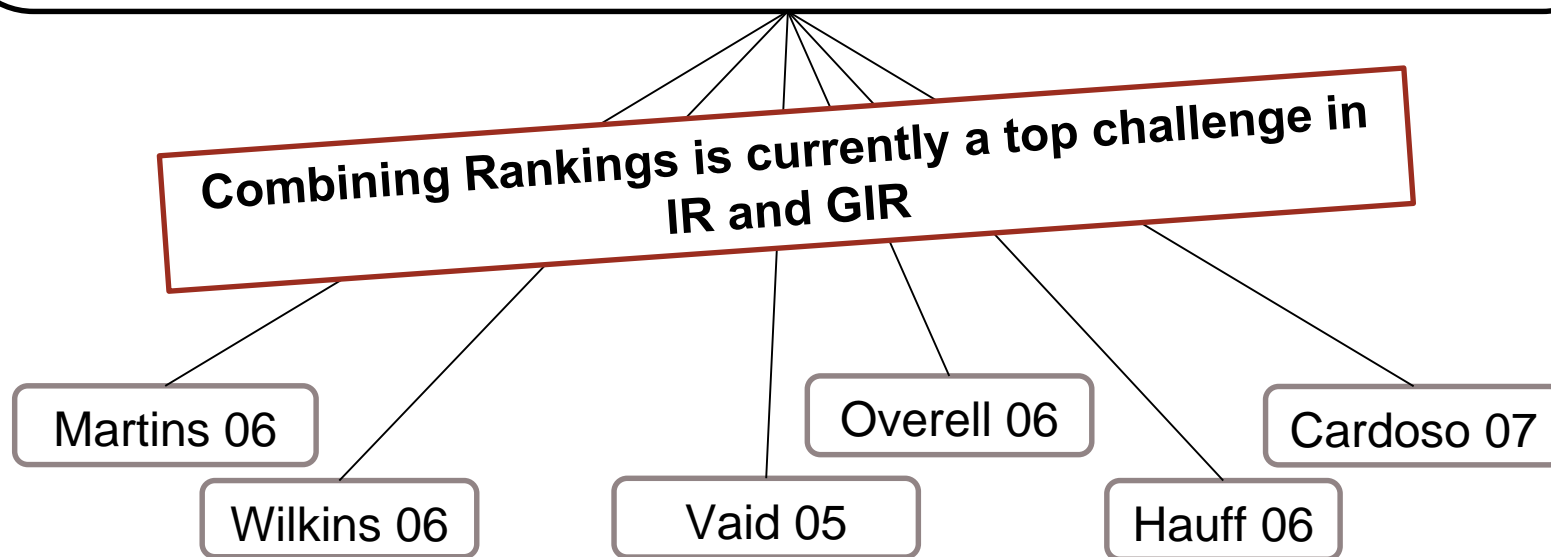
where w_g and w_h are harmonic coefficients of LS and CS respectively, ranging from 0 to 1.

*When **geo-spatial propositions are too complicated**, the qualitative reasoning method cannot assess reasoning belief between arbitrary geospatial propositions.*

The propositions should thus be converted into geo-spatial extent with coordinates, and the spatial similarity is calculated based on quantitative methods, such as overlapping area or Euclidean distance.

Combining Geographic and Textual Ranking (1/4)

Generally in IR it is desirable to browse the results to a query in a single ranked list (in the majority of evaluation forums, this is a requirement). Unless the geographic relevance of a document to a query can be assessed independently of the text relevance, this task requires the **combination of geographic and text relevance**.





Combining Geographic and Textual Ranking (2/4)

Martins 06

- It return a single relevance value as the linear combination between geographic and text relevance.
- Text relevance is calculated using the vector space model with the BM25 term weights.
- Geographic relevance is calculated as the convex combination of three normalized measures: horizontal topographic relevance, vertical topographic relevance and distance

Cardoso 07

- They continue Martins' work testing how multiple geographic footprints can be combined.
- They compare three methods of combining relevance scores: the mean, maximum and Boolean score, where Boolean is equivalent to filtering.
- They found the maximum and Boolean methods to be best.



Combining Geographic and Textual Ranking (3/4)

Wilkins 06

- They examine combining scored lists with a weighting dependent on the distribution of the scores within each list.
- Their hypothesis is that cases where the distribution of scores undergo a rapid initial change correlate with methods that perform well.
- Currently this hypothesis has only been applied to image retrieval but it is applicable to all score based data-fusion applications.

Vaid 05

Hauff 06

Overell 06

- An alternative to a combination of ranks or scores is to filter one rank against another.
- A document cut off value is selected for one rank above which documents are considered relevant and below which not-relevant.
- The other rank can then be filtered against these relevant documents.



Combining Geographic and Textual Ranking (4/4)

The most common ways of combining a geographic and text rank is either as a convex combination of **ranks** or **scores**.

The **advantage** of using **ranks** rather than scores is that the distribution of scores produced by different relevance methods may differ greatly. This problem can be mitigated by normalizing the scores.

Using **ranks** rather than scores has the **disadvantage** that information is discarded.



GIR Evaluation

- The experimental evaluation of IR systems is a subject that has received a lot of attention over the past 40 years.
- IR systems are inherently designed to fulfil a user's information need; testing how well this subjective judgement has been fulfilled is not an easy task.
- Cleverdon [24] proposed the Cranfield methodology. It involves a standard triple of a corpus, queries and relevance judgements (C, Q, R) to be provided allowing different IR systems to be compared. The corpus C is a collection of documents, the queries Q a set of requests for information, and the relevance judgements R a set of documents from the collection that fulfil each information request.



Evaluation Forums (1/2)

- Evaluation forums are now becoming the accepted method of evaluating IR systems.
- The Text REtrieval Conference (TREC) laid the foundation for modern evaluation forums.
- All of the current evaluation forums follow the Cranfield model.
- **GeoCLEF** is the Geographic track at the CLEF (Cross Language Evaluation Forum) forum for comparing IR systems augmented with geographic data. It is becoming the standard for evaluating GIR systems.



Evaluation Forums (2/2)

- The **GeoCLEF 2005-08** English corpus consists of approximately 135,000 news articles, taken from the 1995 Glasgow Herald and the 1994 Los Angeles Times (Gey et al. 2006). The total corpus contains approximately 100M words. There are 100 GeoCLEF queries from 2005-08 (25 from each year). These topics are generated by hand by the four organizing groups.



Evaluation Measures (1/4)

- A contingency table gives an overall picture of results but is generally broken down further into measures.
- Cleverdon (1966) measures of **precision** and **recall** attempt to capture the effectiveness of a retrieval system. Both measures rely on the assumption of relevance that there is a binary measure as to whether documents are relevant or not. **Precision** is the **proportion of retrieved documents** that are **relevant**. **Recall** is the **proportion of relevant documents** that are **retrieved**.
- Generally, both **precision** and **recall** have to be taken into account as there is a trade off between the two.
 - ✓ If the threshold at which documents are considered relevant is increased, fewer documents will be retrieved, precision is expected to rise and recall expected to fall.
 - ✓ Conversely, if the threshold at which documents are considered relevant is decreased, more documents will be retrieved, recall will rise and precision fall.



Evaluation Measures (2/4)

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

- There are many ways to combine **precision** and **recall** into a **single measure** that allows the comparison of different IR systems.
- Dependent on the task and evaluation different measures are viewed as more appropriate. Commonly in retrieval tasks a ranked list will be returned as a result of each query, in which case a **ranked-retrieval measure** may be more appropriate. When the retrieved set is **not ranked** it is common to use the **F-measure**. It is the weighted harmonic mean of precision and recall. Traditionally, the F_1 measure is used where both are equally weighted. This is calculated as:

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$



Evaluation Measures (3/4)

- Intuitively it makes sense that some documents will be more relevant to an information need than others.
- Precision at n (**P@N**) is a measure that models how a system is used. It can be assumed that in a real system a user will not trawl through page after page of results looking for a relevant document (iProspect 2006). It is assumed a user will only look at the first n documents (where n is 5, 10, 20 etc...); the precision is calculated after the first n documents.
- **P@N** is limited as a comparator, as it varies significantly with the number of relevant documents and the selected value of n .
- Average precision (**AP**) is a measure that attempts not to penalize systems for setting the relevance required-for-acceptance threshold too high or too low. It relies on systems being able to quantify relevance and rank documents by their relevance. AP is the average of precisions computed at each relevant document rank.



Evaluation Measures (4/4)

- **F-measure**, **P@N** and **AP** all provide a **single per-query effectiveness value** of an IR system. However, it is common in evaluation forums to represent the effectiveness of a system executed across all queries with a single number (making comparing systems as easy as possible).
- The arithmetic mean is the most common method of combining per-query results. The arithmetic mean of the average precision (short mean average precision or **MAP**) is **the major evaluation measure for IR systems that produce a ranked set of results.**



Information Retrieval Frameworks (1/2)

Apache Lucene is a high-performance, full-featured text search indexing and searching library written entirely in Java. Apache Lucene is highly reputed for its performance and scalability, and is vastly used worldwide. Lucene is developed by the Apache Foundation. See <http://lucene.apache.org/> for more information.

Apache Nutch is an open-source search engine implemented in Java, which uses Apache Lucene. It is a very efficient search engine, but lacking some state-of-the-art ranking algorithms, such as Okapi's BM25. One of its key features is the ability to extend its functionalities through the use of self contained software plugins. Nutch is also developed by the Apache Foundation. For more information, see <http://lucene.apache.org/nutch/>



Information Retrieval Frameworks (2/2)

Lemur Toolkit is an open-source toolkit designed to facilitate research in language modeling and information retrieval. Lemur supports a wide range of industrial and research language applications, such as ad-hoc retrieval, site-search, and text mining. Lemur is implemented in C/C++. See <http://www.lemurproject.org/> for more information.

Terrier is a modular platform for the rapid development of large-scale IR applications, providing indexing and retrieval functionalities, developed by the Information Retrieval Research Group of the Department of Computing Sciences of the University of Glasgow. Terrier has various cutting edge features, including parameter-free probabilistic retrieval approaches (such as Divergence from Randomness models), automatic query expansion/re-formulation methodologies, and efficient data compression techniques. Terrier is written in Java. See <http://ir.dcs.gla.ac.uk/terrier/> for more information.



Terrier Framework

- Terrier is open source, and is a comprehensive, flexible and transparent platform for research and experimentation in text retrieval. Research **can easily be carried out** on standard **TREC** and **CLEF** test **collections**.
- Terrier can index large corpora of documents, and provides multiple indexing strategies, such as multi-pass, single-pass and large-scale **MapReduce** indexing. **Real-time** indexing of document streams are also supported via updatable index structures.
- **State-of-the-art retrieval approaches** are provided, such as Divergence From Randomness, BM25F, as well as term dependence proximity models. Support for supervised ranking models via **Learning to Rank** is also built-in.



Terrier Framework

- Terrier is ideal for **performing information retrieval** experiments. It can index and perform batch retrieval experiments for all known TREC test collections. Tools to evaluate experiments results are also included.
- Terrier uses UTF internally, and can support corpora written in **languages other than English**.
- Terrier follows a **plugin architecture**, and it is easy to extend to develop new retrieval techniques, add new ranking features or experiment with low-level functionality such as index compression.



Some current challenges

- Improve GIR components and methods
 - Geo-tagging, geospatial-textual indexing and geo-relevance ranking.
 - Improved understanding of spatial natural language terminology.
 - Evaluation of GIR systems.
- Integration of geographical with temporal aspects
- Creation of reliable place ontologies with world-wide coverage



References

- [1]** Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008
- [2]** Mark, David M. "Geographic information science: Defining the field." Foundations of geographic information science 1 (2003): 3-18.
- [3]** Yinghua Zhou, Xing Xie, Chuang Wang, Yuchang Gong and Wei-Ying Ma, Hybrid Index Structures for Location-based Web Search, *CIKM'05*, Bremen, Germany, 2005
- [4]** R. Hariharan, B. Hore, C. Li, and S. Mehrotra. Processing spatial lkeyword (sk) queries in geographic information retrieval (gir) systems. In *SSDBM*, page 16, 2007



References

- [5]** A. G. Woodruff, C. Plaunt, Gipsy: Georeferenced information processing system, *Journal of the American Society for Information Science* 45 (9) (1994) 645-655.
- [6]** J. Ding, L. Gravano, N. Shivakumar, Computing geographical scopes of web resources, In *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB '00)*, San Francisco, CA, USA, (2000) 545-556.
- [7]** B. Martins, M. J. Silva, A graph-ranking algorithm for georeferencing documents, in: 2013 IEEE 13th International Conference on Data Mining, IEEE Computer Society, 2005, pp. 741-744.
- [8]** E. Amitay, N. Har'El, R. Sivan, A. Soer, Web-a-where: geotagging web content, in: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2004, pp. 273-280.



References

- [9]** J. L. Leidner, Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names, PhD thesis, Institute of Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.
- [10]** C. E. C. Campelo, C. de Souza Baptista, A model for geographic knowledge extraction on web documents, in: Advances in Conceptual Modeling-Challenging Perspectives, Springer, 2009, pp. 317-326.
- [11]** N. Cardoso, M. J. Silva, Query expansion through geographical feature types, in: Proceedings of the 4th ACM workshop on Geographical information retrieval, ACM, 2007, pp. 55-60.
- [12]** D. Buscaldi, P. Rosso, E. S. Arnal, Using the wordnet ontology in the geoclef geographical information retrieval task, in: GeoCLEF2005, Springer-Verlag, 2006, pp. 939-946.



References

- [13] N. Stokes, Y. Li, A. Moat, J. Rong, An empirical study of the effects of nlp components on geographic ir performance, *International Journal of Geographical Information Science* 22 (3) (2008) 247-264.
- [14] P. Clough, Extracting metadata for spatially-aware information retrieval on the internet, in: *Proceedings of the 2005 workshop on Geographic information retrieval*, ACM, 2005, pp. 25-30.
- [15] M. F. Goodchild, Geographical data modeling, *Computers & Geosciences* 18 (4) (1992) 401-408.
- [16] Larson, R.R. and Frontiera, P., 2004, Spatial ranking methods for geographic information retrieval (GIR) in digital libraries. In *Research and Advanced Technology for Digital Libraries: 8th European Conference, ECDL 2004*, R. Heery and L. Lyon (Eds). September 2004, Bath, UK, *Lecture Notes in Computer Science* 3232 (Berlin: Springer), pp. 45-57.



References

- [17]** Geoffrey Andogah. Geographically Constrained Information Retrieval. PhD thesis, University of Groningen, 2010
- [18]** L. L. Hill. Access to Geographic Concepts in Online Bibliographic Files: Effectiveness of Current Practices and the Potential of a Graphic Interface. PhD thesis, University of Pittsburgh, 1990.
- [19]** Walker, D., I. Newman, D. Medyckyj-Scott and C. Ruggles (1992). "A system for identifying datasets for GIS users." International Journal of Geographical Information Systems 6(6): 511-527
- [20]** M.K. Beard and V. Sharma. Multidimensional Ranking in Digital Spatial Libraries. In Special Issue of Metadata. Journal of Digital Libraries, Vol.1, No. 1, 1997.



References

- [21] Maron, Melvin Earl; KUHNS, John L. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 1960, vol. 7, no 3, p. 216-244.
- [22] Gao, Yong, et al. A Qualitative Representation and Similarity Measurement Method in Geographic Information Retrieval. *arXiv preprint arXiv:1311.4644*, 2013.
- [23] Mark, D. M. (1999). Spatial representation: a cognitive view. *Geographical information systems: principles and applications*, 1, 81-89.
- [24] Cleverdon, C. W. (1970). *The effect of variations in relevance assessments in comparative experimental tests of index languages*. Cranfield, UK: Cranfield Institute of Technology. (Cranfield Library Report No. 3)