Desiderata for Research in Web Intelligence, Mining and Semantics

Rajendra Akerkar
Vestlandsforsking
Sogndal
Norway
rak@vestforsk.no

Costin Bădică
Software Engineering Department
University of Craiova
Craiova, Romania
cbadica@software.ucv.ro

Dumitru Dan Burdescu
Software Engineering Department
University of Craiova
Craiova, Romania
dburdescu@yahoo.com

ABSTRACT
The Web has an immense impact on our daily activities at work, home, and leisure. As a result, more effective and efficient methods and technologies are needed to make the most of the Web's practically unlimited potential. The new Web-related research directions include intelligent methods usually associated with the areas of Computational Intelligence, Semantic Web, Soft Computing, and Data Mining. In this article, the necessity for research on Web intelligence, mining and semantics (WIMS) is discussed together with ways in which a wide range of research is benefiting this area for the long-term. Also the WIMS conference's goal and structure are presented.

Categories and Subject Descriptors
I.2.11 [Artificial Intelligence]: Distributed Artificial Intelligence - intelligent agents

Keywords
Web intelligence, Web semantics, Web mining, big data information retrieval, social networks

1. INTRODUCTION
The rapid growth of Web technology has made the World Wide Web an important and popular application platform for disseminating and searching information as well as for conducting business. This growth gave a way to the development of ever smarter approaches to extract patterns and build knowledge with the aid of artificial intelligence techniques. These techniques have been used, together with information technology, in a wide range of applications. This is where semantics, social network analysis, web structure, content, usage, and other aspects have already been and will increasingly keep being included in many application domains. The Web provides rich medium for communication, which goes far beyond the conventional communication media.

Two characteristics of the Web make it a useful and inimitable platform for computer applications and research, the size and complexity. WIMS community advocate for a new conference series from 2011 devoted to Web Intelligence, Web Mining and Web Semantics. The conference is an international forum for researchers and practitioners to present the state-of-the-art in the area of creating an intelligent Web, to examine performance characteristics of various approaches in Web-based intelligent information technology, and to cross-fertilize ideas on the development of Web-based intelligent information management across different domains.

For the Web to reach its full potential, we must enhance its services, make it more comprehensible, and increase its usability. As researchers continue to develop intelligent tools and techniques for web mining and web semantics, we believe this technology will play ever more vital role in meeting the challenges of developing the intelligent Web.

WIMS'12, June 13-15, 2012, Craiova, Romania.

2. OVERVIEW OF THE RESEARCH FIELDS

2.1 Web Intelligence
Web Intelligence [1] consists of a multidisciplinary area dealing with exploitation of data and services over the Web, to create new data and services using both Information and Communication Technologies (ICT) and Artificial Intelligence (AI) techniques. Different objectives have been followed, and different approaches and technologies have been used by researchers and practitioners over the years. We can mention concepts such as Web information repositories, Web user behavior analysis, Web content and structure mining, social network analysis, the Semantic Web. In addition more general concepts such as Knowledge Discovery from Databases, Multi-Agent Systems, Machine Learning, Knowledge Representation, and Distributed Systems are some keys to understand the fundamentals of intelligent Web.

In particular, the Web has several familiar explicitly and implicitly defined communities. Explicit communities are those that are available to be identified easily on the Web. Kumar et al. [15] discussed the example of an explicit community of Web users interested in Porsche Boxster cars, such as the Porsche newsgroup, or resource collections in directories in search engines, such as the Yahoo directory. Explicit communities are easy to be identified and analysts can merely use manual methods to find an enterprise's explicit communities by browsing the enterprise's newsgroup, or the category in which the enterprise falls into in the directory like Yahoo on the Internet.

Implicit communities are comparatively more complex to find using manual browsing methods. According to Kumar, implicit communities refer to the distributed, ad-hoc and random content-

creation related to the common interests on the Internet [15]. The pages often have links to each other, but the common interests of implicit communities are sometimes too narrow and detailed for the resource pages or the directories to develop explicit listings for them. As a result, it is more difficult to find the implicit communities of an enterprise. In identifying the explicit and implicit communities, it is often assumed that the content pages created by these communities would provide hypertext links back to the enterprise's homepage for reference [16].

With respect to knowledge representation and repositories, areas such as logic, ontology, and reasoning are vital in order to support the basic structure evolving from a Web of data to a Web of knowledge [2]. Besides, once knowledge is mined from the Web data, different standards have been developed to store and manage the different patterns extracted from the content. These repositories have been developed for use in Multidimensional Analysis architectures. This is where Extraction, Transformation, and Loading from Web-based resources, Data Web-based Meta-data Modelling, OLAP queries, and its visualization have been broadly studied [3]. Various Web mining applications, such as Web User Behaviour, Content of Web Sites, and the Analysis of the Web as a graph have been considered in the field of Web Intelligence, Web Mining, Machine Learning, Information Retrieval, and Artificial Intelligence communities in international conferences and journals.

The Web usage mining researchers have extensively investigated Web usability and usefulness considerations, for instance, helping the Web user to obtain information [2]. Moreover, the content of a given Web site has formed the focus for Web Content Mining researchers [1]. The structure, representation, and its analysis have been considered as part of Web structure mining [4] and the information retrieval [5].

In order to offer Web data in suitable formats, Web logs, the Web-site contents, and the Hyperlink Structure of the Web, have been considered as the main source of information. Privacy issues, such as using invasive tools to identify the users, and social network analysis where the user's contacts are exposed, have been the focus of further developments in privacy preserving intelligent web applications [6, 7].

Early research on web structure has led to various ranking algorithms that are now used in the analysis on how communities are formed. This includes the HITS algorithm, where authorities and hubs are established [11]. The web content is being developed by its users in web blogging, virtual communities, online knowledge communities, web forums, microblogging, online encyclopedias, and social network applications. This enables the storage and generation of linked and structured information, that can be associated with text messages and multimedia information such as images and videos. Mining into these various multimedia contents provides insights on users and contributes to determine user profiles.

Recently acclaimed research areas in the field of web intelligence and communities are the social networks and in web communities' analysis [9, 10]. Furthermore, it is also focused on the growth of the semantic Web. The key intention is to give a Web of eloquent meaning. There are various aspects of knowledge representation such as computational linguistics, which have contributed to its development [12]. Several standards for meta-data processing such as the Resource Description Framework (RDF), Web Ontology Language (OWL), semantically interlinked online communities (SIOC), and social network representations of RDF, such as Friend of a Friend (FOAF), have been accepted as the standards to semantics concerns in the Web.

Socially enabled Web information [13] search is a new phenomenon facilitated by recent Web technologies. This collaborative social search involves finding specific people in your network who have the knowledge you're looking for or finding relevant information based on one's social network. People in social groups can provide solutions, pointers to databases or other people (metaknowledge), validation and legitimation of ideas, can serve as memory aids and help with problem reformulation. In [14], a Conversational Search and Recommendation system that involves finding relevant information based on social interactions and feedback along with augmented agent based recommendations is introduced.

2.2 Web Mining

Web mining is the application of data mining techniques to the content, structure, and usage of Web resources. Web mining can profit from given structure on data (as in database tables), but it can also be applied to semi-structured or unstructured data like free-form text. This means that Web mining is a precious help in the transformation from human-understandable content to machine-understandable semantics.

This section exemplifies research problems that must be solved if we are to use data mining techniques efficiently in developing Web intelligence.

The Web itself has been considered from two sides, the structure of the Web as a graph and the semantics of the Web. Research on Web structures investigates several structural properties of graphs arising from the Web, including the graph of hyperlinks, and the graph induced by connections between distributed searches. The study of the Web as a graph is not only interesting in its own right, but also yields valuable insight into Web algorithms for crawling, searching and community discovery, and the sociological phenomena which characterize its evolution [17]. Studies of the semantics of the Web were initiated by Tim Berners-Lee, the creator of the World Wide Web [18]. The Web is referred to as the "Semantic Web", where information will be machine-processable in ways that support intelligent network services such as information/search agents. The Semantic Web requires interoperability standards that address not only the syntactic form of documents but also their semantic content. Semantics also lets agents utilize all the data available on all Web pages, allowing them to gain knowledge from one site and apply it to logical mappings on other sites for ontology-based Web retrieval and e-business intelligence. Ontologies and agent technology can play a fundamental role in facilitating such Web-based knowledge processing, sharing, and reuse between applications.

Mechanically identifying authoritative Web pages for a certain topic is boosting a Web search's quality. The hyperlinks contain an enormous amount of latent human annotation that can help automatically infer the notion of authority. When a Web page's author creates a hyperlink pointing to another Web page, this action can be considered as an approval of that page. The collective endorsement of a given page by different authors on the Web can indicate the importance of the page and lead naturally to

the discovery of authoritative Web pages. So, the Web's linkage data provides a vivid Web mining source.

An index-based Web search engine crawls the Web, indexes Web pages, and builds and stores huge keyword-based indices that help locate sets of Web pages that contain given keywords. By using a collection of closely constrained keywords and phrases, a skilled user can instantly locate relevant documents. However, current keyword-based search engines still suffer from a number of deficiencies. Firstly, a topic of any breadth can easily contain hundreds of thousands of documents. Secondly, several very relevant documents may not have keywords that explicitly define the topic, a phenomenon known as the *polysemy* problem.

To provide potential solutions to these problems, data mining and semantic technologies should be integrated with the Web search engine service to boost the quality of Web searches. Mechanized extraction of Web page structures and semantic contents can be complex given the existing limits on automated natural-language parsing. But, semiautomatic methods can recognize a large segment of such structures. We may nonetheless need to indicate what kinds of structures and semantic contents a specific page type can have. Then a page-structure-extraction system can analyze the Web page to see whether and how a segment's content fit into one of the structures.

In Web mining domain, mining Web log records can uncover association patterns, sequential patterns, and Web access trends. Web access pattern mining often requires taking further measures to obtain additional user traversal information. Such data that can include user browsing sequences from the Web server's buffer pages along with associated data facilitates thorough Web log analysis. Researchers have used these Web log files to analyze system performance, to improve system design through Web caching and page prefetching and swapping, to determine the nature of Web traffic, and to evaluate user reaction to site design. For example, research has been done on adaptive Web sites that improve themselves by learning from user access patterns.

Web log analysis can also help build customized Web services for individual users. Since Web log data provides information about specific pages' popularity and the methods used to access them, this information can be integrated with Web content and linkage structure mining to help rank Web pages, classify Web documents, and construct a multilayered Web information base.

Several data mining methods can help achieve effective Web intelligence. Customizing service to an individual requires tracing that person's Web traversal history to build a profile, then supplying intelligent, personalized Web services based on that information. Nowadays, many Web-based e-commerce service systems, such as amazon.com and tripadvisor.com, register user's previous traversal or purchase history and build customer profiles from that data. Based on a user's profile and preferences, these sites select proper sales promotions and recommendations, thus providing superior service than sites that do not track and store this information. Using data mining to find a user's purchase or traversal patterns can further enhance these services. Although a personalized Web service based on a user's traversal history could help recommend appropriate services, a system usually cannot collect enough information about a particular individual to warrant a quality recommendation. Either the traversal history has too little historical information about that person, or the possible spectrum of recommendations is too broad to set up a history for any one individual. For instance, many people make only a single book purchase, thus providing less data to generate a trustworthy pattern. Here, collaborative filtering is effective because it does not rely on a specific individual's earlier experience but on the collective recommendations of the people who share patterns similar to the individual being examined. Without a doubt, collective filtering has been utilized as a data mining technique for Web intelligence.

Inherently, data mining for Web intelligence will be a key research drive in Web technology. Conversely, we must conquer imminent research challenges before we can make the Web a more intelligent resource that we can all share and explore.

In future, we trust that, Web mining methods will increasingly treat content, structure, and usage in an integrated fashion in iterated cycles of *extracting* and *exploiting* semantics, to be able to comprehend the Web.

2.3 Web Semantics

The Semantic Web plays a crucial role in the development of information technologies and services on the World Wide Web. It takes on new challenges in which the meaning of information enables computers to understand the Web content and imitate human intelligence in performing more of the tedious tasks involved in finding, sharing, and combining information on the Web. Until now computers have not been able to fully accomplish these tasks without human intervention since Web pages are designed to be understood by people, not machines.

The Semantic Web offers a good basis to enrich Web Mining. The types of (hyper) links are now described explicitly, allowing the knowledge engineer to gain deeper insights in Web structure mining; and the contents of the pages come along with a formal semantics, allowing her to apply mining techniques which require more structured input.

Regardless of these initiatives, some of the challenges remain in a bottleneck due to the requirement for automating reasoning systems to deal with inconsistency, vastness, uncertainty, vagueness, and treachery in order to deliver on the promise of the Semantic Web. The discipline of Soft Computing has an evolving collection of methodologies, which aims to exploit tolerance for imprecision, uncertainty, and partial truth to study very complex phenomena: those for which more conventional methods have not yielded low cost and complete solutions. Nowadays, Soft Computing provides a desirable opening for developing Web intelligence to represent the ambiguity in human thinking with real world uncertainty, reason on vagueness in ontologies, and makes possible the transition between the Web and its semantic successor. In this context, the Semantic Web will enable the emergence of digital ecosystems of software and services delivered by the Internet. It will also extend the Internet with capabilities to reason on its resources and their relationships in order to develop the knowledge-based economy in the 21st century.

In recent years an approach to the Semantic Web, called linked data, has been developed that offers a promising path to practical Semantic Web. It provides a set of design guidelines or patterns for how the Semantic Web technologies, and broader Web architecture, can be used for sharing information. Linked Data is a set of conventions for publishing data on the Semantic Web. It is based on principles outlined by Tim Berners-Lee [19]. These principle advocate the use of http URIs for naming entities, the

publication of data about these URIs using the standards (RDF, SPARQL) and inclusion of links to other URIs so that agents can discover more information. While quite simple these guidelines, along with a growing body of practical advice[1], have led to publication and linking of many datasets in this form[2]. However, the existing guidelines and practices have no provision for representation of uncertainty; yet linked data is indeed fraught with many of these different types of uncertainty.

The issue of uncertainty on the Semantic Web is a stimulating research field, as this domain deals with imprecise information from different applications, each with its special knowledge representation needs (e.g., multimedia processing, pattern recognition, etc).

Visualization of Web structure and contents has been another active area of research since the creation of the Web. There are numerous systems for the static visualization and analysis of the link structure of the Web.

In recent days, big data is the buzz word. It is the confluence of three technology trends:

- *Big transaction data*: Massive growth of transaction data volumes
- *Big interaction data*: Explosion of new types of data such as social media and device data
- *Big data processing:* Highly scalable processing with Hadoop

An example of the increase of the relationship between big data and the Semantic Web is Google. In the beginning, Google search eschewed explicit use of semantics, preferring to infer a variety of signals in order to generate results. They used big data to create signals such as PageRank. Now, as the search algorithms mature, search engine companies' mission is to make their results ever more useful to users. Every new data source is a new business opportunity. Whether it's social media data posted by your Facebook fans, device-generated data like call detail records, or the enterprise applications of a recently acquired company, our ability to harness this information bears straight on your bottom line.

Semantic Web systems generate metadata and identified entities explicitly, *i.e.* by hand or as the output of database values. But as anybody who's tried to get users to do it will tell you, generating metadata is hard. This is part of why the full Semantic Web dream isn't yet realized. Analytical approaches take a different angle: surfacing and classifying the metadata from analysis of the actual content and data itself.

Once big data techniques have been effectively applied, we have identified entities and the connections between them. If we want to join that information up to the rest of the Web, or to concepts outside of our system, we need a language in which to do that. We need to organize, to exchange and to reason about those entities. It is the framework that has been gradually developed over the last Decade with the semantic web project.

[1] http://linkeddata.org/

[2] http://linkeddata.org/data-sets

The data landscape is fertile with opportunities to improve performance across multiple domains, yet riddled with the pitfalls posed by rising data volumes, complexity, diversity, and velocity. By harnessing and combining large-scale transactional data with new interaction data and taking advantage of data-intensive frameworks, researchers and data scientists can help organizations leverage their resources to realize the big opportunities of big data and to become a data-centric enterprise.

3. SESSIONS PLANNED AT WIMS'12

In order to meet the challenges of Web related research in the new information age, a new high-impact international conference series, namely the International Conference on Web Intelligence, Mining and Semantics (WIMS) is initiated by Vestlandsforsking in 2011. WIMS'12 is the 2nd meeting in this new series concerned with intelligent approaches to transform the World Wide Web into a global reasoning and semantics-driven computing machine. It is an international forum for researchers and practitioners to present the state-of-the-art in the development of Web intelligence, to examine performance characteristics of various approaches in Web-based intelligent information technology, and to cross-fertilize ideas on the development of Web-based intelligent information systems among different domains. By idea-sharing and discussions on the underlying foundations and the enabling technologies of Web intelligence, we hope to stimulate future development of new models, new methodologies, and new tools for building a variety of embodiments of Intelligent Web Information Systems.

WIMS'12 was hosted by the Software Engineering Department, University of Craiova, Romania during June, 13-15, 2012.

WIMS'12 brought to the research community new results and applications in several topical areas of Web intelligence, mining, and semantics. The conference featured 2 keynotes, 5 tutorials, 8 oral sessions, and 2 poster sessions

The first key note presented by Dr. Elena Simperl addressed methods and challenges for semantic data management enabled by new collective approaches.

The second key note presented by Dr. Jeff Z. Pan addressed the problem of closed world reasoning in Semantic Web.

Tutorial papers introduced to the audience modern methods and technologies for creation and management of semantic content and meta-content, for ontology alignment, for knowledge and reasoning interoperability, as well as for text stream processing.

The papers orally presented at WIMS'12 sessions addressed the following topics: social networks; text and data mining (2 sessions); reasoning, semantics, and ontologies (2 sessions); natural language processing; intelligent agents; information retrieval; linked open data and collaborative systems; and Web applications in e-business and e-learning. The two poster sessions presented research results and applications in various areas of WIMS.

The title "WIMS" of the conference was chosen to reflect the distinct feature that the conference is focused on intelligence and semantic aspects of Web and Web information management and systems.

4. REFERENCES

[1] Akerkar, R. and Lingras, P. 2008. *Building an Intelligent Web: Theory & Practice*. Jones and Bartlett, Sudbury, MA.

[2] Velasquez, J.D., Palade, V.: *Adaptive Web Sites: A Knowledge Extraction from Web Data Approach*. IOS Press, Amsterdam (2008)

[3] Rebolledo, V.L., Velásquez, J.D.: A platform for extracting and storing web data. In: Velásquez, J.D., Rios, S.A., Howlett, R.J., Jain, L.C. (eds.) Knowledge-Based and Intelligent Information and Engineering Systems. LNCS (LNAI), vol. 5711, pp. 843–850.

[4] Liu, B. 2007. *Web Data Mining: Exploring Hyperlinks, Content and Usage Data*, 1st edn. Springer, Heidelberg.

[5] Baeza-Yates, R.A., Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston.

[6] Agrawal, R., Srikant, R. 2000. Privacy-preserving data mining. SIGMOD Rec. 29(2), 439–450.

[7] Xu, Y., Wang, K., Zhang, B., Chen, Z. 2007. Privacy-enhancing personalized web search. In: WWW 2007: Proceedings of the 16th international conference on World Wide Web, pp. 591–600. ACM Press, New York

[8] Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011.

[9] Golbeck, J., Rothstein, M. 2008. Linking social networks on the web with foaf: a semantic web case study. In: AAAI 2008: Proceedings of the 23rd national conference on Artificial intelligence, pp. 1138–1143. AAAI Press, Menlo Park.

[10] Akerkar, R. and Aaberge, T. 2011. Semantically linking virtual communities. (Eds. Christo El Morr & Pierre Maret) Virtual Community Building and the Information Society: Current and Future Directions, pp. 192-207, IGI Global Publishers.

[11] Kleinberg, J.M. 1999. Authoritative sources in a hyperlinked environment. J. ACM 46(5), 604–632.

[12] Shadbolt, N., Berners-Lee, T., Hall, W. 2006. The semantic web revisited. IEEE Intelligent Systems 21(3), 96–101.

[13] Horowitz, D. and Kamvar, S.D. 2010. The Anatomy of a Large-Scale Social Search Engine. Proceedings of ACM WWW 2010.

[14] Venkatesh, A.,Sahay, S., Ram, A. 2010. *Cobot: Real Time Multi User Conversational Search and Recommendations*, Recommender Systems and The Social Web at ACM RecSys

[15] Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. 1998. Trawling the Web for Emerging Cyber-communities, *Proceedings of the 8th International World Wide Web Conference*.

[16] Reid, E. O. F. 2003. Identifying a Company's Non-Customer Online Communities: a Proto-typology. Proceedings of the Hawaii International Conference on System Sciences, Big Island, Hawaii.

[17] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J.L. 2000. Graph structure in the web, Computer Networks, 33, 309-320.

[18] Berners-Lee, T., Hendler, J., and Lassila, O. 2001. The semantic Web, Scientific American, 29-37.

[19] Berners-Lee, T.: Linked Data. 2006. http://www.w3.org/DesignIssues/LinkedData.html