Semantic Web Mining and the Representation, Analysis, and Evolution of Web Space

Bettina Berendt[1], Andreas Hotho[2], and Gerd Stumme[2]

[1] Institute of Information Systems, Humboldt University Berlin, D-10178 Berlin, Germany, http://www.wiwi.hu-berlin.de/~berendt
[2] Knowledge and Data Engineering Group, University of Kassel, D-34121 Kassel, Germany, http://www.kde.cs.uni-kassel.de/[hotho|stumme]

Abstract. Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining. This survey analyzes the convergence of trends from both areas: Growing numbers of researchers work on improving the results of Web Mining by exploiting semantic structures in the Web, and they use Web Mining techniques for building the Semantic Web. Last but not least, these techniques can be used for mining the Semantic Web itself. The second aim of this paper is to use these concepts to circumscribe what Web space is, what it represents and how it can be represented and analyzed. This is used to sketch the role that Semantic Web Mining and the software agents and human agents involved in it can play in the evolution of Web space.

1 Introduction

Faced with the ever-growing importance of the Web, users expect intelligent processing (such as search engines that recognize their true information needs) and a broad and accurate coverage of all realms of their lives (such as information on "real-world" topics or the possibility to utilize services online).

This implies a number of demands on "Web space", the space created by objects that occur in the Web—words, sentences, whole Web pages, hyperlinks, and the traces left behind by Web users. The first user expectation formulated above amounts to a good processing / analysis of Web space; the second to a good representation of the "real world" (or users' views of it): real-life objects that are described in the Web, or represented in Web space, including material things but also more abstract concepts like diseases (in medicine), skills (in education), or any other notion that is the object of human discourse. A good representation in this sense is partly a result of a good analysis of Web space.

At present, three main approaches for reaching these goals can be distinguished: *Web Mining* focuses mostly on the power of statistics for an efficient analysis of the huge amounts of data on the Web and the extraction of knowledge from these data. The *Semantic Web* relies on the potential of explicit knowledge representation to make the Web machine-understandable and support more intelligent processing. (Expressed differently, the Semantic Web aims at Web space

being a formal representation of the worlds that are usually represented informally in today's Web.) We proposed the term *Semantic Web Mining* to denote the combination of these two approaches: The use of semantics to improve mining, and the use of mining to create semantics and thus further the evolution of Web space. This includes, of course, the mining of the Semantic Web itself.

When we organized the first Semantic Web Mining workshop in 2001 [91], there were already a large number of algorithms, tools, and application studies that demonstrated the usefulness of this combination; and their number has risen steadily. Many papers presented at the current workshop, "Representation and Analysis of Web Space", present methods or applications that can be subsumed under Semantic Web Mining.

Therefore, the first aim of this paper is to give a brief overview of important trends in Semantic Web Mining. This is the topic of Sections 2–4, which are based on [13, 10].

The second aim of the paper, and the topic of Section 5, is to consider some future directions that are particularly relevant for an improved understanding and design of Web space, and to show their relations to Semantic Web Mining.

The first direction is a (partial) shift of activity and control: the transition from the Web as a document store to the Web as a collaboration environment in which everyone contributes their knowledge. The current blogs boom as well as large-scale collaborative undertakings like the Wikipedia are just two encouraging signs that this vision may not be as utopic any more as it used to be.

The second is a shift of interest. Apart from the above-mentioned perception of the Web as a tool to handle individual, social, business, and academic life, there is increasingly also the desire to understand the medium Web itself, its dynamics, and its impact on "real life" (such as the self-censorship of dominant search engines in response to political pressure). One may argue that this emerging interest only reflects the Web's becoming a part of real life itself and thus a confluence of Web space and what it represents. Nonetheless, this view brings with it many new formal and semantic challenges.

2 Semantic Web Mining: Goals and Foundations

Semantic Web Mining aims at combining the two areas Semantic Web and Web Mining by using semantics to improve mining and using mining to create semantics. Last but not least, these techniques can be used for mining the Semantic Web itself.

In the past few years, there have been many attempts at "breaking the syntax barrier" on the Web. A number of them rely on the semantic information in text corpora that is implicitly exploited by statistical methods. Some methods also analyze the structural characteristics of data; they profit from standardized syntax like XML. In this paper, we concentrate on markup and mining approaches that refer to an *explicit conceptualization* of entities in the respective domain. These relate the syntactic tokens to background knowledge represented

in a model with *formal semantics*. When we use the term "semantic", we thus have in mind a formal logical model to represent knowledge.

2.1 The analysis of Web space: Web data and Web Mining

Web mining is the application of data mining techniques to the content, structure, and usage of Web resources. Three areas of Web mining are commonly distinguished: content mining, structure mining, and usage mining [98, 62, 87]. In all three areas, a wide range of general data mining techniques, such as association rule discovery, clustering, classification, and sequence mining, are employed and developed further to reflect the specific structures of Web data and the Web-related application questions. Typical tasks include document classification and clustering, data mining for information extraction, network analysis for document ranking, usage mining for dynamic personalization, system adaptation, or modelling networks of users. Recent overviews of Web mining methods and applications are given in [10, 87]; recent textbooks include [7, 19].

Learning methods construct models from samples of structured data. Most methods are defined for samples in which each case is defined as values of a fixed set of features. In Web content mining (which is usually text mining), documents are cases, and a feature-value representation of a document must be constructed. A standard approach is to treat words as features and word occurrences as values. This requires a number of preprocessing steps like tokenization, stemming, the removal of stopwords and formatting elements, synonym merging, etc. In addition, other features of documents can be used (for example, length, occurrence of numbers, number of images).

The hyperlinks induce a graph structure on the set of Web pages. Based on the observation that a hyperlink, like a citation in scientific work, usually represents a "vote of confidence", this hyperlink structure is used to identify pages of high quality. (Citations of the type "The book/Web page XYZ is totally wrong" are the exception and are generally ignored.) One example is the PageRank algorithm [16] that is implemented in the Google search engine. It defines the importance of a Web page in a site based on the number of links from other important sites. HITS [59] follows a similar scheme, but differentiates between two types of pages. An "authority" is a page which is pointed to from many important hubs, and "hubs" are pages pointing to many important authorities.

Depending on the application and the questions of analysis, different data are recorded in user logs and analyzed in usage mining. At the lowest level clicks on menu items and keystrokes can be recorded. At a higher level, commands, queries, entered text, drawings can be logged. The context in which user actions are observed can be the entire screen, a menu or other. The context can include textual documents. The content and usage of the Web can be viewed as single units but also as structures. The content consists of pieces that are connected to other pieces in several ways: by hyperlinks (possibly with labels), addresses, textual references, shared topics or shared users. Similarly, users are related by hyperlinks, electronic or postal addresses, shared documents, pages or sites.

Web usage mining is characterized by the need of an extensive data preparation. Web server data are often incomplete, missing a unique association between a user and her activities and a complete record of her activities.

One of the aims of the Semantic Web, discussed next, is to avoid the manifold problems of Web data interpretation by separating presentation from content.

2.2 The Semantic Web: Ontologies for representing (our views of) the world

Ontologies are at the core of the well-known layer structure for the Semantic Web (`http://www.w3.org/DesignIssues/Semantic.html`), providing the opportunity of representing arbitrary worlds. An ontology is "an explicit formalization of a shared understanding of a conceptualization" [40]. This high-level definition is realized differently by different research communities, but most definitions include a set of *concepts*, a hierarchy on them, and (n-ary) *relations* between concepts. The relations may be linked to one another by a relation hierarchy. Most of them also include axioms in some specific logic [90, 15]. Depending on whether one needs axioms or not, one can use OWL or RDF Schema to formalize ontologies that conform to the definition.

Ontologies are formed of concepts; Web pages usually describe concrete instances of these concepts in human-readable form. Metadata are the intermediates between these two representations; their objects (identified by URIs) can be seen as instances of the ontology concepts.

Tools for creating, learning, and maintaining ontologies and for building ontology-based applications include Protégé-2000 [79], KAON [15], and Sesame [54]. An extensive overview of ontology tools can be found in [39].

Besides the formal languages for the Semantic Web, ontologies for general use are developed. At present, there are mainly in practice two types of ontologies. The first type uses a small number of relations between concepts, usually the subclass relation and sometimes the part-of relation. Popular and commonly used are ontologies of Web documents, such as DMoz or Yahoo!, where the documents are hierarchically organized based on content (for example: "Computers" – "Data Formats" – "Markup Languages"). For each content topic (such as "XML" below "Markup Languages"), there is an ontology node, and this is associated with (usually several hundreds of) Web pages identified by their URLs.

The other kind of ontologies are rich with relations but have a rather limited description of concepts, usually consisting of a short definition. A well-known example of a general, manually constructed ontology is the semantic network WordNet (`http://wordnet.princeton.edu`) with 26 different relations (e.g., hypernym, synonym). For instance, the concept "bird" is connected to "animal" via "is a kind of" and to "wing" via "has part".

3 Extracting Semantics from the Web

Web Mining can help to learn structures for knowledge organization (e. g., ontologies) and to provide the population of such knowledge structures.

It should be noted that all approaches discussed here are semi-automatic, assisting the knowledge engineer in her task. In order to obtain high-quality results, one cannot replace the human in the loop, as there is always a lot of tacit knowledge involved in the modeling process [18]. A computer will never be able to fully consider background knowledge, experience, or social conventions. If this were the case, the Semantic Web would be superfluous, since then machines like search engines or agents could operate directly on conventional Web pages. The overall aim of our research is thus not to replace the human, but rather to provide her with more and more support.

Ontology Learning Extracting an ontology from the Web is a challenging task; engineering the ontology by hand is expensive. In [70], the expression *Ontology Learning* was coined for the semi-automatic extraction of semantics from the Web. There, machine learning techniques were used to improve the ontology engineering process and to reduce the effort for the knowledge engineer.

Ontology learning exploits many existing resources including texts, thesauri, dictionaries, and databases. It builds on techniques from Web content mining, and it combines machine learning techniques with methods from fields like information retrieval [68] and agents [95], applying them to discover the 'semantics' in the data and to make them explicit. The techniques produce intermediate results which must finally be integrated in a machine-understandable format, e. g., an ontology.

Ontology learning is related to the fully automated extraction of relations like hypernymity/hyponymity from text (see [37] for a Web-scale implementation that takes advantage of the massive redundancy of the Web and employs machine learning for choosing which patterns to accept). Again, from the point of view of Semantic Web Mining, the obtained patterns need to be checked for correctness in order to integrate them into an ontology under construction.

A particular challenge lies in the enrichment of an existing ontology with explicit semantics like WordNet by statistical information that is given semantic meaning [78]. Mining can supplement taxonomies with new categories (cf. [3] for an extension of Yahoo!), and it can help build new taxonomies [61].

Mapping and Merging Ontologies The growing use of ontologies leads to overlaps between knowledge in a common domain. Domain-specific ontologies are modeled by multiple authors in multiple settings. These ontologies lay the foundation for building new domain-specific ontologies in similar domains by assembling and extending multiple ontologies from repositories.

The process of *ontology merging* takes as input two (or more) source ontologies and returns a merged ontology. Manual ontology merging using conventional editing tools without support is difficult, labor-intensive, and error-prone. Therefore, several systems and frameworks for supporting the knowledge engineer in the ontology merging task have recently been proposed. Overviews of approaches and tools are given in [80, 81]. These approaches rely on syntactic and semantic

matching heuristics which are derived from the behavior of ontology engineers confronted with the task of merging ontologies.

Ontology mapping is the assignment of the concepts of one ontology and their instances to the concepts of another ontology. This could be useful, for example, when one of several ontologies has been chosen as the right one for the task at hand. The instances can simply be classified from scratch into the target ontology; alternatively, the knowledge inherent in the source ontology can be utilized by relying on the heuristic that instances from one source concept are likely to also be classified together in one concept of the target ontology [99].

An alternative to merging/mapping ontologies is to collect a 'corpus of representations' [41] and to select the right one according to the task at hand. In Section 5, some of the issues arising from the availability of multiple ontologies are discussed further.

Instance Learning Even if ontologies are present and users manually annotate new documents, there will still be old documents containing unstructured material. In general, the manual markup of every produced document is impossible. Also, some users may need to extract and use different or additional information from the one provided by the creator. To build the Semantic Web, it is therefore essential to produce automatic or semi-automatic methods for extracting information from Web-related documents as instances of concepts from an ontology, either for helping authors to annotate new documents or for extracting additional information from existing unstructured or partially structured documents.

A number of studies investigate the use of content mining to enrich existing conceptualizations behind a Web site. For example, text categorization techniques can be used to assign HTML pages to categories in the Yahoo hierarchy [77]. This can reduce the manual effort for maintaining the Yahoo Web index.

Information Extraction from texts (IE) is one of the most promising areas of Natural Language Technologies (see, e. g., [31]). IE is a set of automatic methods for locating important facts in electronic documents for subsequent use. IE techniques range from the extraction of keywords from pages' text using the *tf.idf* method known from Information Retrieval, via techniques that take the syntactic structures of HTML or natural language into account, to techniques that extract with reference to an explicitly modeled target structure such as an ontology (for a survey, see [65]). Interactive techniques that help human annotators by highlighting etc. are a good supplement to IE for document analysis and related tasks, e.g., [48, 32, 43]. In [46, 47], machine learning techniques were used for the semi-automatic annotation of web services.

Using existing conceptualizations as ontologies and for automatic annotation For many sites, an explicit domain model for the generation of web pages already exists as a database or in a Content Management System. These existing formalizations can be (re-)used for semantic markup and mining, e.g. [12]. In [44], "deep annotation" derives mappings between information struc-

tures. These mappings are used for querying semantic information stored in the database underlying the web site.

Semantics created by structure As we have discussed in Section 2.1, the results of the analysis of Web page linkage by Web usage mining create a ranking of relevance. Another kind of knowledge that may be inferred from structure is a similarity between pages, useful for the popular browser application "Find similar pages" (to one that has been retrieved by browsing or search); see [33] for the use of co-citation analysis, the observation that pages which are frequently cited together from other pages are often related. These techniques structure the set of pages, but they do not classify them into an ontology.

In contrast, the hyperlink structure within pages lends itself more directly to classification. An ontology of page functions ("head", "navigation", and "content" pages), and a classification of pages into these categories based on their textual and hyperlink content, is proposed in [30, 84]. The structure of within-page markup may also help in extracting page content: concentrating on page segments identified by reference to the page's DOM (document object model, or tag tree) can serve to identify the main content of a page [19, pp. 228ff.] and to separate it from "noise" like navigation bars, advertisements, etc. [96].

Semantics created by Usage The preceding discussion has implicitly assumed that content exist independently of its usage. However, a large proportion of knowledge is socially constructed. Thus, navigation is not only driven by formalized relationships or the underlying logic of the available Web resources. Rather, it "is an information browsing strategy that takes advantage of the behavior of like-minded people" ([22, p.18]). Collaborative filtering and related approaches serve to learn relations of "relatedness" between items ("People who liked/bought this book also looked at ..."). How can this be extended to learning more semantically specific information?

Navigational traces may be used to learn content categories. In [97], navigation is described by hidden Markov models, with the hidden states being page categories, and the observed request events being instances of them. Semantic labels (such as "sports pages") must be assigned to a state manually.

Based on frequent paths through a site and the keywords extracted from the pages along the path, the intended goal or "information scent" that was followed can be identified [24, 23]. The information scent is a set of weighted keywords, which can be inspected and labeled more concisely using an interactive tool. Thus, usage creates a set of information goals users expect the site to satisfy (for an empirical validation, see [25]).

Keywords that occur in pages or along paths can also be processed further to extract hidden semantic commonalities (latent semantic indexing and related techniques, cf. [55] for an application to usage mining); the identification of semantics as defined in this paper then requires further processing.

User interest in a site's concepts is measured by the frequency of accesses to pages that deal with these concepts in [88]. They use these data for *ontology*

evolution: extending the site's coverage of high-interest concepts, and deleting low-interest concepts, or merging them with others.

The combination of implicit user input (usage) and explicit user input (search engine queries) can contribute further to conceptual structure. User navigation has been employed to infer topical relatedness, i.e. the relatedness of a set of pages to a topic as given by the terms of a query to a search engine ("collaborative crawling" [1]). A classification of pages into "satisfying the user defined predicate" and "not satisfying the predicate" is thus learned from usage, structure, and content information. An obvious application is to mine user navigation to improve search engine ranking [56, 58].

Many approaches use a combination of content and usage mining to generate recommendations. For example, in content-based collaborative filtering, textual categorization of documents is used for generating pseudo-rankings for every user-document pair [75]. In [83], ontologies, IE techniques for analyzing single pages, and a user's search history together serve to derive recommendations for query improvement in a search engine.

4 Using Semantics for Web Mining and Mining the Semantic Web

Semantics can be utilized for Web Mining for different purposes. Some of the approaches presented in this section rely on a comparatively *ad hoc* formalization of semantics, while others can already exploit the full power of the Semantic Web. 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. Because the distinction between the use of semantics for Web mining and the mining of the Semantic Web itself is all but sharp, we will discuss both in an integrated fashion.

Content and Structure Mining In [50], ontologies are used as background knowledge during preprocessing, with the aim of improving clustering results. We preprocess the input data (e. g., text) and apply ontology-based heuristics for feature selection and feature aggregation. Resulting improvements in clustering and classification, using WordNet as the background ontology, are described in [52, 14].

In [78], the general-purpose WordNet is used to improve ontology learning from text in specialized domains, and the learned concepts are added as refining subtrees to that base ontology.

Ontology-based search hierarchies of concepts and multiple search paths, together with conceptual clustering, are used to facilitate the customized access to email [29] and courseware material that is stored in a peer-to-peer network (`http://edutella.jxta.org`). A combination of this approach based on Formal Concept Analysis with text clustering is presented in [51].

Knowledge-rich approaches in automatic text summarization (cf. [72, 73, 53]) aim at maximizing the information within a minimal amount of text. They are closely related to web content mining using semantics because in both Web content mining and text summarization, natural language text needs to be mapped into an abstract representation. This abstract is often represented in some logic, and it is used to improve the results of text summarization. We expect that techniques for automatic text summarization will play an important role in Semantic Web Mining.

Web structure mining can also be improved by taking content into account. The PageRank algorithm mentioned in Section 2.1 co-operates with a keyword analysis algorithm, but the two are independent of one another. So PageRank will consider any much-cited page as 'relevant', regardless of whether that page's content reflects the query. By also taking the hyperlink anchor text and its surroundings into account, CLEVER [20] can more specifically assess the relevance for a given query. The Focused Crawler [21] improves on this by integrating topical content into the link graph model, and by a more flexible way of crawling. The learning Intelligent Crawler [2] extends the Focused Crawler, allowing predicates that combine different kinds of topical queries, keyword queries, or other constraints on the pages' content or meta-information (e.g., URL domain). Ontology-based focused crawling is proposed by [69].

Usage Mining Web usage mining benefits from including semantics into the mining process for the simple reason that the application expert as the end user of mining results is interested in *events in the application domain*, in particular user behavior, while the data available—Web server logs—are technically oriented sequences of *HTTP requests*. The key idea lies in formulating ontologies of atomic application events (AAEs), specifying how page requests instantiate one or more AAEs, and formalizing complex application events (CAEs), usually as sequences or regular expressions of AAEs. The occurrence of atomic or complex AE patterns in the data is then detected using mappings from usage data to AEs. A detailed formalization and literature overview are given in [11].

In the analysis and evaluation of user behavior, it must be kept in mind that different stakeholders have different perspectives on the usage of a site, which leads them to investigate different processes (complex application events) and also makes them consider different user actions 'correct' or 'valuable'. Recently, frameworks have been proposed for capturing different processes [67, 92, 4] and perspectives [76].

Closing the Loop One of the key promises of Semantic Web Mining is that all newly gained knowledge about Web space is represented formally. This means that this knowledge can be fed back into the human-machine system, presenting the next user (or the same user next time) with a richer representation of the world and of Web space.

In [13], we have described one out of many possible combinations of the approaches presented in the previous sections. The goal is to take a set of Web

pages from a site and to improve them for both human and machine users: (a) to generate metadata that reflect a semantic model underlying the site, (b) to identify patterns both in the pages' text and in their usage, and, based on these insights, to improve information architecture and page design. To achieve these goals, we have proposed a series of steps in which we

- employ mining methods on Web resources to generate semantic structure (learning and filling the ontology),
- employ mining methods on the resulting semantically structured Web resources to generate further structure,
- at the end of each step, feed these results back into the content and design of the Web pages themselves (visible to human users) and/or their metadata and the underlying ontology (visible to machine users).

5 Semantic Web Minng in Context: An Outlook

Web mining and mass collaboration Semantics can be created by learning from people's externalized knowledge (texts, structure between texts, and behaviour), but first, there are many areas that still defy this approach and that require the elicitation of contextualized human input as a starting point, and second, combining information obtained from documents and obtained by asking may greatly improve quality. Examples include the labelling of non-textual material such as pictures for purposes such as multimedia retrieval, paraphrasing or translating natural language for purposes such as question answering, or collecting commonsense knowledge. A paid staff of specially-qualified experts is expensive, and manual input is error-prone if done by only one person.

The phenomenal success of mass collaboration of volunteer contributors in such diverse areas as open source software (e.g., Linux), ontology creation (e.g., the Open Directory Project), or joint authorship (e.g., Wikipedia) has spawned a number of large-scale projects that collect knowledge and/or metadata. A particularly interesting observation is a recent confluence of methodology between mining-based Web-scale knowledge collection on the one hand and mass-collaboration-based efforts on the other.

(1) Ontology learning from large text corpora rests on simple templates, often based on Hearst's [45] seminal work on extracting hyponym/hyperonym relations ("cities *such as* London, Paris, and Tokyo", "collies, dachshunds, *and other* dogs", ...), e.g., [37]. As linguists point out, there are many other natural-language patterns for extracting interesting relations between concepts, propositional attitudes, etc. [93]. However, template-based learning from texts may fail for certain relations (reported for meronymity in [45]) or for certain domains ("is a" is too generic for technical domains and must be replaced by other templates better suited to a corpus of definitions, [78]). Similarly, systems that learn from volunteer contributors' input generally ask directly for input to templates (e.g., ontology learning in nontechnical domains [89], commonsense knowledge [86], paraphrases in a restricted commonsense domain [27], or knowledge in a technical domain [85]). In [35], users are enlisted for a wide range of tasks to support

information integration: They provide training data and domain knowledge, and they verify tool predictions.

(2) Quality control has to remove inconsistencies and errors and deal with multiple occurrences of the same piece of knowledge. To this end, statistical as well as symbolic inference mechanisms are employed: pointwise mutual information [37, 93], equality of answers between two or more anonymous contributors (which is statistically extremely rare in a large search space) [94, 27], Bayesian inference aggregation [85], and various forms of propositional/predicate inference [89, 86]. A manual evaluation of commonsense knowledge collected from volunteers [28] suggests that measures similar to association rules' high support and low confidence are indicators of high-quality contributions.

An interesting alternative to count-based evaluations of contributions are reputation systems and trust networks, which are fed by system-generated tests [35] or by explicit user ratings (`www.bibserv.org`, see [85]). This mechanism can be said to resemble focused crawling in mining-based semantics creation. As [82] point out, the issue of trust arises on all levels, including the trust in or usefulness of ontologies. They propose a system to obtain this "second-level Semantic Web knowledge" from volunteers that uses many of the ideas mentioned above, but also point out that an infinite regress (who will rate the raters) must be avoided.

Incentives for participation and good user interfaces are crucial for getting volunteers' contributions, and they are extensively discussed in the articles in [26]. Games give very good incentives, e.g., [94, 27]; in domains that are inherently less "fun" such as Web search or scientific writing, domain-related useful payoffs are needed [35, 8]. In addition, if knowledge collection and use are to become truly global and open for everyone, the impact of individual differences must be considered. This includes the impact of language and culture on usage [64, 63] and on other motivational factors.

Indeed, mass collaboration may be required to address a phenomenon that belies the very idea of the Semantic Web: "Even today (and the situation will get only worse), it is often easier to develop a new ontology from scratch than to reuse someone else's ontology that is already available" [82, p. 56], an observation that is also supported by the generally extremely low ontology ranks (number of references from other ontologies in the repository) displayed in the search engine Swoogle (`swoogle.umbc.edu`, [34]).

Web structure and evolution – towards a representation of Web space
While the Semantic Web's primary focus was on better (i.e. more machine-understandable) representations of "the real world", the past few years have witnessed a rising interest in the Web itself and the dynamics of this medium. Most analyses have focused on structural characteristics of Web subgraphs and their implications. For example, Web Communities are defined by being a graph clique, but are also considered to identify systems of semantically interconnected content, user interests, and users [38], and their evolution changes Web space.

If one wants to model Web space at this more semantic level that also captures the meaning of second-level entities like Web Communities, the first challenge

consists of defining an ontology to describe these entities, their attributes, and the relations between them. One of the most encompassing proposals in this field has been made by [17]. They distinguish a MAIN component of the Web (the sites that are in the strongly connected component of the connectivity graph of sites), IN (sites that can reach MAIN but cannot be reached from MAIN), OUT (sites that can be reached from MAIN but not vice versa), TENTACLEs of IN or OUT, TUNNELs between IN and OUT, and ISLANDs (unconnected sites). This classification is based on structure like the page classification described in Section 3, and like that one, it is linked to function (for example, typical ISLANDs are eCommerce sites that apparently try to avoid linking to competitors).

This model has been subjected to a large-scale empirical evaluation with respect to the Chilean Web [5, 6]. The site classification was extended slightly, and the components were investigated with respect to features derived from the sites and the pages that constitute the components, in particular age- and freshness-related features ("the kernel of the Web comes mainly from the past" [5, p. 8]), the relation between the structural features that define the components and the average hubness and authority of their sites. In addition, movements of sites between the components have been described, showing for example that the main growth of the Web is in the ISLANDS component, but that this component also has the largest incidence of site death, followed by OUT. On the other hand, MAIN is quite stable but growing over the years, which may indicate a more mature Web. It grows mainly from OUT or NEW sites.

This study focusses on the description of states and changes, and the authors speculate that "perhaps there are no formal processes behind [these dynamics] and [the current movements are] just a transient phase" [6, p. 7]. This presumption could be reconsidered in the light of recent studies that model and simulate different citation / linkage "cultures" and show their effects on network structure (see `www.kinf.wiai.uni-bamberg.de/COM` and [71]), demonstrating that there is substantial scope for formal and semantic models of citation, reference, and their impact on Web space and its dynamics.

Summary and Outlook In this paper, we have studied Semantic Web Mining, the combination of the two fast-developing research areas Semantic Web and Web Mining. We expect that in the future, Web mining methods will increasingly treat content, structure, and usage in an integrated fashion in iterated cycles of *extracting* and *utilizing* semantics, to be able to understand and (re)shape the Web. Among those iterated cycles, we expect to see a productive complementarity between those relying on semantics in the sense of the Semantic Web, and those that rely on a looser notion of semantics.

While the latter have the advantage that they can operate without human intervention, at Web scale, and (sometimes) in real time, approaches that involve explicit semantics and human quality control are needed in high-commitment domains that require correct and exhaustive knowledge (such as science or business). In addition, an explicit conceptualization enables people and programs to explain, reason, and argue about meaning and thus rationalize their trust, or

lack of trust, in a system. Therefore, Semantic Web Mining is the best support for the development of principled feedback loops that consolidates the knowledge extracted by mining into information available for the Web at large. In addition, it will enable us to integrate results from machine learning and mass collaboration, and to reason about the newly-emerging objects in today's highly dynamic Web spaces.

References

1. C.C. Aggarwal. Collaborative crawling: Mining user experiences for topical resource discovery. In *[42]*, pages 423–428, 2002.
2. C.C. Aggarwal, F. Al-Garawi, and P.S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. In *Proceedings of the WWW Conference*, 2001.
3. C.C. Aggarwal, S.C. Gates, and P.S. Yu. On the merits of building categorization systems by supervised clustering. In *KDD'1999 – Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 352–356, 1999.
4. S.S. Anand, M. Mulvenna, and K. Chevalier. On the deployment of web usage mining. In *[9]*, pages 23–42. 2004.
5. Ricardo Baeza-Yates, Carlos Castillo, and Felipe Saint-Jean. Web dynamics, structure and page quality. In M. Levene and A. Poulovassilis, editors, *Web Dynamics*, pages 93–109. Springer, 2004.
6. Ricardo A. Baeza-Yates and Barbara Poblete. Dynamics of the chilean web structure. In *Proceedings of the 3rd International Workshop on Web Dynamics at WWW 2004. New York, 18th May 2004*, 2003.
7. P. Baldi, P. Frasconi, and P. Smyth, editors. *Modeling the Internet and the Web. Probabilistic Methods and Algorithms*. Wiley, Chichester, UK, 2003.
8. B. Berendt. Understanding and supporting volunteer contributors: The case of metadata and document servers. In *[26]*, pages 106–109. 2005.
9. B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, and G. Stumme, editors. *Web Mining: From Web to Semantic Web. First European Web Mining Forum, EWMF 2003. Invited and Selected Revised Papers*, volume 3209 of *LNAI*. Springer, Berlin, 2004.
10. B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou and, and G. Stumme. A roadmap for web mining: From web to semantic web. In *[9]*, pages 1–22. 2004.
11. B. Berendt, A. Hotho, and G. Stumme. Usage mining for and on the semantic web. In *[57]*, pages 461–480. 2004.
12. B. Berendt and M. Spiliopoulou. Analys of navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal*, 9(1):56–75, 2000.
13. Bettina Berendt, Andreas Hotho, and Gerd Stumme. Towards semantic web mining. In *International Semantic Web Conference*, pages 264–278, 2002.
14. Stephan Bloehdorn and Andreas Hotho. Text classification by boosting weak learners based on terms and concepts. In *Proceedings of the Fourth IEEE International Conference on Data Mining*. IEEE Computer Society Press, 2004.
15. E. Bozsak, M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, G. Stumme, Y. Sure, J. Tane, R. Volz, and V. Zacharias. Kaon - towards a large scale semantic web. In K. Bauknecht, A. Min Tjoa, and G. Quirchmayr, editors, *E-Commerce and Web Technologies, Third International Conference, EC-Web 2002, Proceedings*, volume 2455 of *LNCS*, pages 304–313, Berlin, 2002. Springer.
16. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *WWW7 / Computer Networks*, 30(1-7):107–117, 1998.
17. Andrei Z. Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet L. Wiener. Graph structure in the web. *WWW9 / Computer Networks*, 33(1-6):309–320, 2000.
18. B. Buchanan. Informed knowledge discovery: Using prior knowledge in discovery programs. In *KDD 2000 – Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, August 20-23, 2000*, page 3, New York, 2000. ACM.
19. S. Chakrabarti. *mining the web*. Morgan Kaufmann, San Francisco, CA, 2003.
20. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th World-wide web conference (WWW7),30(1-7)*, pages 65–74, 1998.
21. S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. *Computer Networks*, 31:1623–1640, 1999.
22. C. Chen. *Information Visualisation and Virtual Environments*. Springer, London, 1999.

23. E.H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions on the web. In *Proceedings of the ACM CHI 2001 Conference on Human Factors in Computing Systems*, pages 490–497, Amsterdam: ACM Press, 2001.
24. E.H. Chi, P. Pirolli, and J. Pitkow. The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a web site. In *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*, pages 161–168, Amsterdam: ACM Press., 2000.
25. E.H. Chi, A. Rosien, and J. Heer. Intelligent discovery and analysis of web user traffic composition. In *[74]*, pages 1–15, 2002.
26. T. Chklovski, P. Domingos, H. Lieberman, R. Mihalcea, and P. Singh, editors. *Knowledge Collection from Volunteer Contributors. Papers from the AAAI 2005 Symposium*, volume SS-05-03 of *Technical Report*. Menlo Park, CA: AAAI Press, 2005.
27. Timothy Chklovski. 1001 paraphrases: Incenting responsible contributions in collecting paraphrases from volunteers. In *[26]*, pages 16–20. 2005.
28. Timothy Chklovski and Yolanda Gil. Towards managing knowledge collection from volunteer contributors. In *[26]*, pages 21–27. 2005.
29. R. Cole and G. Stumme. Cem - a conceptual email manager. In B. Ganter and G. W. Mineau, editors, *Proc. ICCS 2000*, volume 1867 of *LNAI*, pages 438–452. Springer, 2000.
30. R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems*, 1(1):5–32, 1999.
31. J. Cowie and Y. Wilks. Information extraction. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*. Marcel Dekker, New York, 2000.
32. M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, 118(1-2):69–113, 2000.
33. J. Dean and M.R. Henzinger. Finding related pages in the world wide web. In *Proceedings of the Eighth International World Wide Web Conference WWW-1999*, Toronto, May 1999.
34. L. Ding, T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Riddivari, V. Doshi, and J. Sachs. Swoogle: a search and metadata engine for the semantic web. In *Proc. of the 13th ACM Conference on Information and Knowledge Management*, pages 652–659, 2004.
35. A. Doan, R. McCann, and W. Shen. Collaborative development of information integration systems. In *[26]*, pages 34–41. 2005.
36. P. Domingos, C. Faloutsos, T. Senator, H. Kargupta, and L. Getoor, editors. *KDD'2003 – Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 2003. ACM.
37. O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates. Methods for domain-independent information extraction from the web: An experimental comparison. In *Proc. of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pages 391–398, Menlo Park, CA, 2004. AAAI/MIT Press.
38. G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of Web communities. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, 2000. ACM Press.
39. Asun Gomez-Perez, Juergen Angele, Mariano Fernandez-Lopez, V. Christophides, Athur Stutt, and York Sure. A survey on ontology tools. OntoWeb deliverable 1.3, Universidad Politecnia de Madrid, 2002.
40. T. R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, Netherlands, 1993. Kluwer.
41. Alon Y. Halevy and Jayant Madhavan. Corpus-based knowledge representation. In Georg Gottlob and Toby Walsh, editors, *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, pages 1567–1572. Morgan Kaufmann, 2003.
42. D. Hand, D. Keim, and R. Ng, editors. *KDD - 2002 – Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 2002. ACM.
43. Siegfried Handschuh and Steffen Staab. Authoring and annotation of web pages in CREAM. In *Proc. Of WWW11*. to appear, 2002.
44. Siegfried Handschuh, Steffen Staab, and Raphael Volz. On deep annotation. In *Proc. of WWW-2003*, Budapest, Hungary, 05 2003.
45. Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992, 14th International Conference on Computational Linguistics, August 23-28, 1992, Nantes, France.*, pages 539–545, 1992.
46. Andreas Heß, Eddie Johnston, and Nicholas Kushmerick. ASSAM: A tool for semi-automatically annotating web services with semantic metadata. In *Proc. Intl. Semantic Web Conference (ISWC 2004)*. Springer, 2004 (to appear).
47. Andreas Hess and Nicholas Kushmerick. Learning to attach semantic metadata to web services. In *The Semantic Web – Proc. Intl. Semantic Web Conference (ISWC 2003)*, pages 258–273. Springer, 2003.

48. Jerry Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson. Fastus: A cascaded finite-state transducer for extracting information from natural-language text. In E. Roche and Y. Schabes, editors, *Finite State Devices for Natural Language Processing*. MIT Press, Cambridge MA, 1996.
49. I. Horrocks and J. A. Hendler, editors. *The Semantic Web – ISWC 2002, First International Semantic Web Conference, Proceedings*, volume 2342 of *LNCS*. Springer, 2002.
50. A. Hotho, A. Maedche, and S. Staab. Ontology-based text clustering. In *Proceedings of the IJCAI-2001 Workshop "Text Learning: Beyond Supervision", August, Seattle, USA*, 2001.
51. A. Hotho, S. Staab, and G. Stumme. Explaining text clustering results using semantic structures. In *[66]*, pages 217–228, 2003.
52. A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Proc. of the ICDM 03, The 2003 IEEE International Conference on Data Mining*, pages 541–544, 2003.
53. M. Hu and B. Liu. Mining and summarizing customer reviews. In *[60]*, pages 695–700, 2004.
54. Frank van Harmelen Jeen Broekstra, Arjohn Kampman. Sesame: A generic architecture for storing and querying rdf and rdf schema. In *[49]*, pages 54–68, 2002.
55. X. Jin, Y. Zhou, and B. Mobasher. Web usage mining based on probabilistic latent semantic analysis. In *[60]*, pages 197–205, 2004.
56. T. Joachims. Optimizing search engines using clickthrough data. In *[42]*, pages 133–142, 2002.
57. H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, editors. *Data Mining: Next Generation Challenges and Future Directions*. AAAI/MIT Press, Menlo Park, CA, 2004.
58. C. Kemp and K. Ramamohanarao. Long-term learning for web search engines. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2002)*, pages 263–274, Berlin, 2002. Springer.
59. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
60. R. Kohavi, J. Gehrke, W. DuMouchel, and J. Ghosh, editors. *KDD'2004 – Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 2004. ACM.
61. D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proceedings of the 14th International Conference on Machine Learning (ML), Nashville, Tennessee, July 1997*, pages 170–178, 1997.
62. R. Kosala and H. Blockeel. Web mining research: A survey. *SIGKDD Explorations*, 2(1), 2000.
63. A. Kralisch and B. Berendt. Language-sensitive search behaviour and the role of domain knowledge. submitted.
64. A. Kralisch and B. Berendt. Cultural determinants of search behaviour on websites. In *Proceedings of the IWIPS 2004 Conference on Culture, Trust, and Design Innovation*, pages 61–74, Vancouver, BC, 2004. Product & Systems Internationalisation, Inc.
65. A.H.F. Laender, B.A. Ribeiro-Neto, A.S. da Silva, and J.S. Teixeira. A brief survey of web data extraction tools. *SIGMOD Record*, 31(2):84–93, 2002.
66. N. Lavrač, D. Gamberger, L. Todorovski, and H. Blockeel, editors. *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases: PKDD 2003*, volume 2838 of *LNAI*, Berlin Heidelberg, 2003. Springer.
67. J. Lee, M. Podlaseck, E. Schonberg, and R. Hoch. Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Data Mining and Knowledge Discovery*, 5(1/2):59–84, 2001.
68. A. Maedche. *Ontology Learning for the Semantic Web*. Kluwer, 2002.
69. A. Maedche, M. Ehrig, S. Handschuh, L. Stojanovic, and R. Volz. Ontology-focused crawling of documents and relational metadata. In *Proceedings of the Eleventh International World Wide Web Conference WWW-2002*, Hawaii, 2002.
70. A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.
71. Thomas Malsch, Christoph Schlieder, Peter Kiefer, Maren Lübcke, Rasco Perschke, Marco Schmitt, and Klaus Stein. Communication between process and structure: Modelling and simulating message-reference-networks with COM/TE. *JASSS*. submitted.
72. I. Mani and M.T. Maybury, editors. *Advances in Automatic Text Summarization*. The MIT Press, 1999.
73. Inderjeet Mani. *Automatic Summarization*, volume 3 of *Natural Language Processing*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2001.
74. B. Masand, M. Spiliopoulou, J. Srivastava, and O.R. Zaïane, editors. *Workshop Notes of the Fourth WEBKDD Web Mining for Usage Patterns & User Profiles at KDD'2002*, Edmonton, Alberta, Canada, July 23 2002. ACM.
75. P. Melville, R.J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering. In *Proceedings of the ACM SIGIR Workshop on Recommender Systems*, Sep 2001.
76. E. Menasalvas, S. Millán, M.S. Pérez, E. Hochsztain, and A. Tasistro. An approach to estimate the value of user sessions using multiple viewpoints and goals. In *[9]*, pages 164–180. 2004.
77. Dunja Mladenic. Turning yahoo to automatic web-page classifier. In *European Conference on Artificial Intelligence*, pages 473–474, 1998.

78. R. Navigli and P. Velardi. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7), 2005.
79. N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, and M. A. Musen. Creating semantic web contents with protg-2000. *IEEE Intelligent Systems*, 16(2):60–71, 2001.
80. Natalya Fridman Noy. Semantic integration: A survey of ontology-based approaches. *SIGMOD Record*, 33(4):65–70, 2004.
81. Natalya Fridman Noy. Tools for mapping and merging ontologies. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 365–384. Springer, 2004.
82. N.F. Noy, R. Guha, and M.A. Musen. User ratings of ontologies: Who will rate the raters? In *[26]*, pages 56–63. 2005.
83. S. Parent, B. Mobasher, and S. Lytinen. An adaptive agent for web exploration based of concept hierarchies. In *Proceedings of the 9th International Conference on Human Computer Interaction*, New Orleans, LA, 2001.
84. Ramana Rao Peter Pirolli, James Pitkow. Silk from a sow's ear: Extracting usable structures from the web. In *Proc. ACM Conf. Human Factors in Computing Systems, CHI*, pages 118–125, New York, NY, 1996. ACM Press.
85. Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. Trust management for the semantic web. In Dieter Fensel, Katia P. Sycara, and John Mylopoulos, editors, *International Semantic Web Conference*, volume 2870 of *Lecture Notes in Computer Science*, pages 351–368. Springer, 2003.
86. Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In Robert Meersman and Zahir Tari, editors, *CoopIS/DOA/ODBASE*, volume 2519 of *Lecture Notes in Computer Science*, pages 1223–1237. Springer, 2002.
87. J. Srivastava, P. Desikan, and V. Kumar. Web mining – concepts, applications, and research directions. In *[57]*, pages 405–423. 2004.
88. N. Stojanovic, A. Maedche, B. Motik, and N. Stojanovic. User-driven ontology evolution management. In *Proceedings of the 13^{th} European Conference on Knowledge Engineering and Knowledge Management EKAW'02*, 2002.
89. D.G. Stork and C.P. Lam. Open mind animals: Insuring the quality of data openly contributed over the world wide web. In *AAAI Workshop on Learning with Imbalanced Data Sets*, pages 4–9, 2000.
90. G. Stumme. Using ontologies and formal concept analysis for organizing business knowledge. In J. Becker and R. Knackstedt, editors, *Wissensmanagement mit Referenzmodellen – Konzepte fr die Anwendungssystem- und Organisationsgestaltung*, pages 163–174, Heidelberg, 2002. Physica.
91. G. Stumme, A. Hotho, and B. Berendt, editors. *Semantic Web Mining*, Freiburg, September 3rd 2001. 12th Europ. Conf. on Machine Learning (ECML'01) / 5th Europ. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'01).
92. M. Teltzrow and B. Berendt. Web-usage-based success metrics for multi-channel businesses. In *Proc. of the WebKDD Workshop on Web Mining and Web Usage Analysis*, pages 17–27, 2003.
93. Lucy Vanderwende. Volunteers created the web. In *[26]*, pages 84–90. 2005.
94. Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In Elizabeth Dykstra-Erickson and Manfred Tscheligi, editors, *CHI*, pages 319–326. ACM, 2004.
95. A.B. Williams and C. Tsatsoulis. An instance-based approach for identifying candidate ontology relations within a multi-agent system. In *Proceedings of the First Workshop on Ontology Learning OL'2000*, Berlin, Germany, 2000. Fourteenth European Conference on Artificial Intelligence.
96. L. Yi, B. Liu, and X. Li. Eliminating noisy information in web pages for data mining. In *[36]*, pages 296–305, 2003.
97. A. Ypma and T. Heskes. Categorization of web pages and user clustering with mixtures of hidden markov models. In *[74]*, pages 31–43, 2002.
98. Osmar R. Zaïane. From resource discovery to knowledge discovery on the internet. Technical Report TR 1998-13, Simon Fraser University, 1998.
99. D. Zhang and W.S. Leee. Learning to integrate web taxonomies. *Journal of Web Semantics*, 2(2):131–151, 2004.