Mining the World Wide Web – Methods, Applications, and Perspectives

Andreas Hotho, Gerd Stumme

"Some people have advocated transforming the Web into a massive layered database to facilitate data mining, but the Web is too dynamic and chaotic to be tamed in this manner." O. Etzoni, 1996 [10]

1 Introduction

The World Wide Web has become, over the last years, a major source of information, and at the same time a significant platform for commerce. Both aspects make it an interesting target for data mining applications. In this survey, we will discuss different facets of data mining on the Web, and illustrate its methods by typical application areas. These areas will be highlighted in a more detailed descriptions in the subsequent contributions to this special issue of the KI Journal on Web Mining. As internet based applications become more and more intertwined, we will equally consider related domains like email and newsgroups here.

The contributions of the special issue indicate new trends in web mining research. Although not specifically requested in the call for papers, they all focus on one of two issues: the detection of upcoming topics and trends, or the detection and support of online communities. We discuss in this paper that the emergence of these application domains goes together with two technical developments: the Semantic Web for explicitly representing knowledge in the Web, and the Web 2.0 as an effort for facilitating user participation in the Web. We will argue that the convergence of these two areas – one being an academic, top-down and the other a grass-roots, bottom-up approach – will be a major research challenge for the next years, where web mining will play a significant role.

2 Web Mining

Web Mining can broadly be seen as the application of adapted data mining methods to the web. Whereas data mining is defined as the application of algorithm to find patters on mostly structured data embedded into a general knowledge discovery process [12] web mining has the special property to provide a set of different data types. The data

Different to 'classical' data, the Web has different facets that yield different approaches for the mining process: (i) web pages consist of text, (ii) web pages are linked via hyperlinks, and (iii) user activity can be monitored via web server logs. These three facets lead to the distinction into the three areas of *web content mining*, *web structure mining*, and *web usage mining*.

Content Mining. For web content mining, each web page is considered as an individual document. Sets of web pages form a document collection, on which text mining techniques can be applied (see [7, 9] for an overview). One can take advantage of the semi-structured nature of web pages, as HTML provides information that concerns not only layout, but also logical structure.

Another typical content mining task is information extraction where structured information is extracted from unstructured web sites. The goal is to facilitate information aggregation over different web sites by using the extracted structured information. Typical applications are price comparison sites or news aggregators [22].

Web content mining can be used to identify topics in the web, as the contributions of Berendt/Draheim, Hoser et al, and Stein/zu Eißen to this volume show. Recommender also make extensive use of content mining techniques, as discussed in this volume by Semeraro et al and Mobasher.

Structure Mining. For web structure mining, one considers the web (or parts of it) as a directed graph, with the web pages (or whole web sites) being the vertices, that are connected by hyperlinks. The most prominent application in this regard is definitely the Google search engine, which computes the ranking of its results primarily with the PageRank algorithm [30]. It defines a page to be highly relevant if frequently linked by other highly relevant pages, see Section 3 for more details.

Structure and content mining approaches are often combined. Some authors subsume both approaches together under the term 'content mining' (as opposed to web usage mining). Examples for such a combination is the work on trend detection in newsgroups by Hoser et al and the work on community evolution of Falkowski/Spiliopoulou in this volume.

Usage Mining. For mining the usage of web resources, one is considering records of requests of visitors of a web site, that are usually collected as web server logs [31]. While content and structure of collections of web pages reflects the intentions of the author(s) of the pages, the user requests indicate how the consumer perceives these pages. Web usage mining may reveal relationships that were not intended by the creator of the pages. A typical application are correlations in buying behavior, that may be used for recommendations ("People who bought x also bought y."); see for instance [28, 25]). Another application is the discovery of frequent navigation sequences [8, 21], which may be used for a re-design of the website. Web usage mining is frequently combined with content and structure analysis for investigating the semantics of the observed navigation patterns.

3 Mining Methods

All of the existing data mining techniques (eg, clustering, classification, association rules [16]) can also be applied in web mining. They usually need a more extensive preprocessing, since web resources are normally not in the form of a flat table which is required for most methods. (For web content mining which is very similar to text mining, for instance, the usual approach is the transformation of the web pages into 'bags of words', usually including stopword removal and stemming [19].) As we assume that these approaches are widely known by now, we focus here on techniques that have a special bias towards the structure of web data.

Link Mining. One spouseless property of the web are the links which are typically represent the structure of the web. Analyzing these links is is often referred as *Link Mining* and is becoming very popular is the last years. Link Mining is mainly divided into three major tasks (cf. [15]): the object related task like clustering based on links, prediction of (missing) links and a graph centered task like subgraph discovery. A good overview is given in [15]. Techniques with a special emphasis on the graph structure are also topic of further development in the area of social network analysis, as discussed below.

Statistical Relation Learning. Statistical Relation Learning (SRL) focuses on the combination of probabilistic and logic models with the goal to develop one combined approach which is better able to describes real world phenomena. The main idea behind this is to overcome the limits of both worlds. Whereas traditional statistical machine learning is able to capture uncertainty but only at one relation traditional ILP and relational learning approaches are able to work on multiple relation but can't handle noise. The combination of both tries to overcome these limitations. Methods developed in this area are typically applied on richly structural data which are available for e.g. Hypertext classification, topic prediction of bibliographic entries, or on any kind of social networks. The upcoming book [14] will provide an introduction into this area.

Social Network Analysis. SNA has a long-standing tradition, with important steps being the modeling of social relationships within 'sociograms' in the 30ies [29], and the application of graph analysis techniques for sociological and anthroprosophical studies from the 60ies on [17]. The name 'Social Network Analysis' (SNA) was coined in the 70ies. SNA techniques analyse the network as a whole, or study properties of the individual nodes. Measures for the network as a whole comprise density (percentage of present edges among possible edges), diameter (length of the longest shortest path between any two nodes), clustering coefficient, etc. In a web mining scenario, these may be used, together with other staticistal measures, in the data understanding phase.

A key notion for studying individual nodes is 'centrality', which comes in different flavors [11]: In/out degree centrality measures the number of in/outbound edges of a node, and may, e. g., be used for analysing the social status of people. Betweenness centrality measures the number of shortest paths a node is lying on; a typical application is within (technical or social) communication networks. A third line of research in SNA combines the properties of individual nodes with (the growth of) the overall network, leading a. o. to a model of 'preferential attachment' [1].

There are several implementations of SNA algorithms and frameworks, the most prominent being Pajek,[1] UCINet,[2] and Visone.[3] Wasserman and Faust [32] and Freeman [13] provide good surveys on SNA.

4 Applications

In order to illustrate the use of the methods and algorithms discussed above, we will now sketch some prototypical application scenarios. Because of the variety of uses of the web, this list is of course far from being complete.

Search. Because of its large size, search engines have become a crucial navigation component of the Web. The most prominent one, Google, uses the PageRank algorithm [5] which computes the eigenvector centrality of the web graph. This measure reflects the idea that the relevance of a node increases when more relevant nodes point to it. The The Hyperlink-Induced Topic Search [23] follows the same approach for topic-specific web search. It additionally distinguishes between *authorities* (which are referred to by many hubs), and *hubs* (which point to many authorities). Other search engines like Vivisimo[4] apply clustering techniques or

Recommendations. Personalised marketing has become an important business issue in the past few years. Based on the easiness of collecting personal information on the web and the possibility of dynamic web page generation, recommender systems have become a major web application. They make extensive use of data mining techniques, as discuss the articles of B. Mobasher and Semeraro et al in this volume.

Topic & Trend Detection. The dynamic nature of the web provides another challenge: new topics may come suddenly, or trends may emerge slowly. Topics and trends can be detected by comparing the evolution of web content, structure and/or usage over time. The articles by Berendt/Degemmis, Hoser et al, and Stein/zu Eißen in this volume present different examples.

Communities. With the rise of the Web 2.0 (see below), social networks and communities have become in the interest of many researchers. Data mining and social network analysis methods have been deployed to discover communities, and to study their evolution over time. Two examples are presented in the contributions of Falkowski/Spiliopoulou and of Hoser et al to this volume.

[1] http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[2] http://www.analytictech.com/ucinet.htm
[3] http://www.visone.info/ [4] http://vivisimo.com/

5 Beyond the Current Web

Two strong trends have influenced the evolution of the Web: Semantic Web and Web 2.0. While the former is more academia driven, the latter arose in a bottom-up manner. We discuss web mining with respect to these two developments, before analysing its importance for the upcoming convergence of both approaches in the outlook.

Semantic Web. The Semantic Web is based on the vision of Tim Berners-Lee, the inventor of the WWW, to enrich the web by machine-processable information to support the user in his tasks. In this line, the Resource Description Framework RDF (modeling edge and vertice labeled graphs using XML syntax) and the Web Ontology Language OWL (combining RDF and RDF Schema with description logics[5]) have been defined.[6] The Semantic Web provides at the same time interesting resources to be mined, and a formalism for representing results of (web) mining.

Extracting Semantics from the Web. A backbone of the Semantic Web are ontologies, which at present are often handcrafted. This is not a scalable solution for a wide-range application of Semantic Web technologies. The challenge is to learn ontologies in a (semi-)automatic way [26]. The ontologies can then be filled using Information Extraction [24] methods.

Exploiting Semantics for Web Mining and Mining the Semantic Web. Background knowledge – in the form of ontologies, or in other forms – can be used to improve the process and results of Web Mining. Recent developments include the mining of sites that become more and more semantically enriched sites (e. g., in [4]), as well as directly mining ontologies (e. g., in [18]).

In [3] we have discussed these perspectives in more detail under the notion of 'Semantic Web Mining'.

Web 2.0 . Social networking systems, blogs, wikis, and social bookmarking tools have rapidly emerged on the Web. One reason for their immediate success is the fact that no specific skills are needed for participating. At the moment, these systems provide only very simple structures for organising knowledge – very much in contrast to the semantic web approach. Nevertheless, more structured approaches become necessary once the systems grow – but they should still be usable by non-experts. A key question is Web 2.0 research is: How will current and emerging Web 2.0 systems support untrained users in sharing knowledge on the Web within the next years? This question has to be addressed by combining approaches from different areas like data and web mining, information retrieval, ontology engineering, natural language processing, social network analysis, library and information science, and hypermedia systems.

The different types of Web 2.0 systems pose different challenges for mining algorithms: (i) Social bookmarking tools like del.icio.us[7] or BibSonomy,[8] for instance, are based on so-called folksonomies, i. e., three-dimensional (user/tag/resource) binary tensors. First steps in adapting mining approaches to this structure have been done, e. g., for triadic frequent closed itemsets [20] or clustering [2]. The evolution of network properties of folksonomies has been studied in [6]. (ii) Social Network systems like XING/openBC[9] provide shortest paths in the social network between oneself and other users one is interested in – a computationally expensive operation. An interesting challenge is also the detection of communities, i. e., of groups of users with similar interests and/or with short network distances. (iii) Blogs contributing a huge mass of of new information in a distributed way. Also a discussion of one topic is carried out over several blogs but within existing communities. To identify such communities and make the provided topic specific information accessible allows for the detection of new topics and trends in such communities. (iv) Wikis can be seen as an unstructured information repository. Challenges are the enhancement of current wikis by more structured representation and the analysis of existing information to convert it into more structured once.

Common to all these systems is a large growth and highly dynamic change of its contents, which makes time an important dimension in the mining process. The detection of trends and topic shifts is thus a favorite research issue in this domain.

6 Outlook

With the Web 2.0 still in its infancy, the notion of 'Web 3.0' was coined at the end of last year [27], which emphasizes on the use of Semantic Web technology in combination with Web 2.0 techniques. Even though this new buzzword probably will do more harm than good, it points in the right direction. We definitely assume that the gap between the bottom-up oriented Web 2.0 approaches and the top-down structured Semantic Web knowledge representation will have to be closed. On one hand, Web 2.0 applications gather large groups of people who are willingly providing content, and on the other hand, there is, with the Semantic Web, a sophisticated formalism for representing knowledge. Classical data and web mining techniques have to be extended on both ends to bridge the gap between this two areas, as they do neither fully take into account the new data structures of the Web 2.0 as input format nor the rich knowledge representation within the Semantic Web for output. Very encouraging perspectives for research in the area of web mining indeed!

Literatur

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[2] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, 2006.

[3] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In *Proc Int. Semantic Web Conference*, Sardinia, Italy, 2002.

[4] S. Bloehdorn and A. Hotho. *Boosting for Text Classification with Semantic Features*. 2006.

[5] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.

[5] http://dl.kr.org/

[6] http://www.w3.org/2001/sw/

[7] http://del.icio.us

[8] http://www.bibsonomy.org

[9] http://www.xing.com/

[6] C. Cattuto, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. Network properties of folksonomies. *AI Communications Special Issue on Network Analysis in Natural Sciences and Engineering"(to appear)*, 2007.

[7] S. Chakrabarti. Data mining for hypertext: A tutorial survey. *SIGKDD Explorations*, 1:1–11, 2000.

[8] R. Cooley. *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data*. PhD thesis, University of Minnesota, Faculty of the Graduate School, 2000.

[9] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In *Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*. IEEE Computer Society, Nov 1997.

[10] O. Etzioni. The world-wide web: Quagmire or gold mine? *Communications of the ACM*, 39(11):65–68, 1996.

[11] M. Everett and S. Borgatti. The centrality of groups and classes. *Journal of Mathematical Sociology*, 23(3):181–201, 1999.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, pages 37–54, 1996.

[13] L. C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. BookSurge Publishing, 2004.

[14] B. T. L. Getoor, editor. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. 2007.

[15] L. Getoor and C. P. Diehl. Link mining: a survey. *SIGKDD Explor. Newsl.*, 7(2):3–12, 2005.

[16] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 1st edition, 2000.

[17] F. Harary, R. Z. Norman, and D. Cartwright. *Structural models : an introduction to the theory of directed graphs*. Wiley, New York, 1965.

[18] B. Hoser, A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Semantic network analysis of ontologies. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 514–529, Heidelberg, June 2006. Springer.

[19] A. Hotho, A. Nürnberger, and G. Paaß. A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20(1):19–62, MAY 2005.

[20] R. Jäschke, A. Hotho, C. Schmitz, B. Ganter, and G. Stumme. Trias - an algorithm for mining iceberg tri-lattices. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 06)*, pages 907–911, Hong Kong, December 2006. IEEE Computer Society.

[21] H. Kato, T. Nakayama, and Y. Yamane. Navigation analysis tool based on the correlation between contents distribution and access patterns. In *Proc. WebKDD 2000*, pages 95–104, Boston, MA, 2000.

[22] M. Kayed and K. F. Shaalan. A survey of web information extraction systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411–1428, 2006. Member-Chia-Hui Chang and Member-Moheb Ramzy Girgis.

[23] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[24] A. Laender, B. Ribeiro-Neto, A. da Silva, and J. Teixeira. A brief survey of web data extraction tools. *SIGMOD Record*, 31(2):84–93, 2002.

[25] W. Lin, S. Alvarez, and C. Ruiz. Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery*, 6:83–105, 2002.

[26] A. Maedche. *Ontology Learning for the Semantic Web*. Kluwer, 2002.

[27] J. Markoff. Entrepreneurs see a web guided by common sense. New York Times, November 12 2006. http://www.nytimes.com/2006/11/12/business/12web.html?ex

[28] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.

[29] J. Moreno. *Who shall survive? : a new approach to the problem of Human Interrelations*, volume 58 of *Nervous and mental disease monograph series*. Nervous and Mental Disease Publ., Washington, 1934.

[30] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[31] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and application of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[32] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*, volume 8 of *Structural Analysis in the Social Sciences*. Cambridge University Press, Cambridge, 1st edition, 1999.

Contact

Dr. Andreas Hotho, Prof. Dr. Gerd Stumme
Hertie Chair on Knowledge and Data Engineering
University of Kassel
Wilhelmshöher Allee 73, 34121 Kassel
Tel.: +49 (0)561-8046250
Email: {hotho, stumme}@cs.uni-kassel.de
http://www.kde.cs.uni-kassel.de/{hotho, stumme}
http://www.bibsonomy.org/group/kde

Bild

Andreas Hotho holds a Ph.D. from the University of Karlsruhe, where he worked from 1999 to 2004 at the Institute AIFB. He earned his Master's Degree in information systems from the University of Braunschweig (Germany) in 1998. Since 2004 he is a senior researcher at the University of Kassel. His focus is on the combination of machine learning/data mining and semantic web and also on the analysis of Web 2.0 data.

Bild

Gerd Stumme is heading the Hertie Chair on Knowledge & Data Engineering at the University of Kassel and member of the Research Center L3S since 2004. He earned his PhD in 1997 at Darmstadt University of Technology, and his Habilitation at the Institute AIFB of the University of Karlsruhe in 2002. In 1999/2000 he was Visiting Professor at the University of Clermont-Ferrand, France, and Substitute Professor for Machine Learning and Knowledge Discovery at the University of Magdeburg in 2003.