Editorial: Special Issue on Web Content Mining

Bing Liu
Department of Computer Science
University of Illinois at Chicago
851 South Morgan Street
Chicago, IL 60607-7053
liub@cs.uic.edu

Kevin Chen-Chuan Chang
Department of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Avenue
Urbana, IL 61801-2302
kcchang@cs.uiuc.edu

1. INTRODUCTION

With the phenomenal growth of the Web, there is an ever-increasing volume of data and information published in numerous Web pages. The research in Web mining aims to develop new techniques to effectively extract and mine useful knowledge or information from these Web pages [8]. Due to the heterogeneity and lack of structure of Web data, automated discovery of targeted or unexpected knowledge/information is a challenging task. It calls for novel methods that draw from a wide range of fields spanning data mining, machine learning, natural language processing, statistics, databases, and information retrieval. In the past few years, there was a rapid expansion of activities in the Web mining field, which consists of Web usage mining, Web structure mining, and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks. Web content mining aims to extract/mine useful information or knowledge from Web page contents. For this special issue, we focus on *Web content mining*.

The objectives of this special issue are two-fold:

1. To bring together and to present some of the latest research results in the field. We are not aware of any dedicated issue in any journal on this important topic.
2. To encourage more research activities in the field. With the huge amount of data/information already on the Web and more to come, the next big thing is naturally how to make best use of the Web to mine useful data/information and to integrate heterogeneous data/information automatically.

In this Editorial, we begin by reviewing some of the important problems in Web content mining. We then introduce the papers published in this issue.

2. TOPICS OF WEB CONTENT MINING

It is often said that the Web offers an unprecedented opportunity and challenge for data mining. We believe that this is so due to the following characteristics of the Web:

1. The amount of data/information on the Web is huge and still growing rapidly. Web data is also easily accessible.
2. The coverage of Web information is wide and diverse. One can find information about almost anything on the Web.
3. Data of all types exist on the Web, e.g., structured tables, texts, multimedia data (e.g., images and movies), etc.
4. Information on the Web is heterogeneous. Multiple Web pages may present the same or similar information using completely different formats or syntaxes, which makes integration of information a challenging task.
5. Much of the Web information is semi-structured due to the nested structure of HTML code, and the need of Web page designers to present information in a simple and regular fashion to facilitate human viewing and browsing.
6. Much of the Web information is linked. There are links among pages within a site, and across different sites. These links serve as an information organization tool and also as indications of trust/authority in the linked pages and sites.
7. Much of the Web information is redundant. The same piece of information or its variations may appear in many pages or sites. This property has been explored in many Web data mining tasks.
8. The Web is noisy. A Web page typically contains a mixture of many kinds of information, e.g., main content, advertisements, navigation panels, copyright notices, etc. For a particular application only part of the information is useful, and the rest are noises.
9. The Web consists of surface Web and deep Web. Surface Web is composed of pages that can be browsed using a normal Web browser. Surface Web is also searchable through popular search engines. Deep Web is mainly composed of databases that can only be accessed through parameterized queries using query forms.
10. The Web is also about services. Many Web sites and pages enable people to perform operations with input parameters, i.e., they provide services.
11. Above all, the Web is a virtual society. It is not only about data, information and services, but also about interactions among people, organizations and automatic systems.
12. The Web is dynamic. Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues for many applications.

We can see why the Web is such a fascinating place and why it offers so many opportunities for data mining. Below, we give an introduction to some of the current Web content mining tasks. We will not discuss traditional mining tasks that are directly applied to the Web data, e.g., Web page classification and Web page clustering.

2.1 Structured Data Extraction

This is perhaps the most widely studied research topic of Web content mining. One of the reasons for its importance and popularity is that structured data on the Web are often very important as they represent their host pages' essential information, e.g., lists of products and services. Extracting such

data allows one to provide value added services, e.g., comparative shopping, and meta-search. Structured data is also easier to extract compared to unstructured texts. This problem has been studied by researchers in AI, database and data mining, and Web communities. There are several approaches to structured data extraction, which is also called wrapper generation. The first approach is to manually write an extraction program for each Web site based on observed format patterns of the site. This approach is very labor intensive and time consuming. It thus does not scale to a large number of sites. The second approach is wrapper induction or wrapper learning, which is the main technique currently. Wrapper learning works as follows: The user first manually labels a set of trained pages. A learning system then generates rules from the training pages. The resulting rules are then applied to extract target items from Web pages. Example wrapper induction systems include WIEN [27], Stalker [39], BWI [18], WL^2 [13], etc. The third approach is the automatic approach. Since structured data objects on the Web are normally database records retrieved from underlying databases and displayed in Web pages with some fixed templates. Automatic methods aim to find patterns/grammars from the Web pages and then use them to extract data. Examples of automatic systems include IEPAD [9], MDR [33], RoadRunner [15], EXALG [3], [19], [31], [42], etc.

2.2 Unstructured Text Extraction

Most Web pages can be seen as text documents. Extracting information from Web documents has also been studied by many researchers. The research is closely related to text mining, information retrieval and natural language processing. Current techniques are mainly based on machine learning and natural language processing to learn extraction rules from manual labeled examples [6, 7, 14, 24, 29, 46]. Recently, a number of researchers also make use of common language patterns (common sentence structures used to express certain facts or relations) and redundancy of information on the Web to find concepts, relations among concepts and named entities [12, 20, 32]. The patterns can be automatically learnt or supplied by human users. Another direction of research in this area is Web question-answering. Although question-answering was first studied in information retrieval literature, it becomes very important on the Web as Web offers the largest source of information and the objectives of many Web search queries are to obtain answers to some simple questions. [28, 43] extend question-answering to the Web by query transformation, query expansion, and then selection.

2.3 Web Information Integration

Due to the sheer scale of the Web and diverse authorships, various Web sites may use different syntaxes to express similar or related information. In order to make use of or to extract information from multiple sites to provide value added services, e.g., metasearch, deep Web search, etc, one needs to semantically integrate information from multiple sources. Recently, several researchers attempted this task. Two popular problems related to the Web are (1) Web query interface integration, to enable querying multiple Web databases (which are hidden in the deep Web) [21, 22, 23, 57, 50], and (2) schema matching, e.g., integrating Yahoo and Google's directories to match concepts in the hierarchies [2, 17]. The ability to query multiple deep Web databases is attractive and interesting because the deep Web contains a huge amount of information or data that is not indexed by general search engines [5, 10].

2.4 Building Concept Hierarchies

Because of the huge size of the Web, organization of information is obviously an important issue. Although it is hard to organize the whole Web, it is feasible to organize Web search results of a given query. A linear list of ranked pages produced by search engines is insufficient for many applications. The standard method for information organization is concept hierarchy and/or categorization. The popular technique for hierarchy construction is text clustering, which groups similar search results together in a hierarchical fashion. Several researchers have attempted the task using clustering [11, 26, 30, 36, 55, 56]. In [32], a different approach is proposed which does not use clustering. Instead, it exploits existing organizational structures in the original Web documents, emphasizing tags and language patterns to perform data mining to find important concepts, sub-concepts and their hierarchical relationships. In order words, it makes use of the information redundancy property and semi-structure nature of the Web to find what concepts are important and what their relationships might be. This work aims to compile a survey article or a book on the Web automatically.

2.5 Segmenting Web Pages & Detecting Noise

A typical Web page consists of many blocks or areas, e.g., main content areas, navigation areas, advertisements, etc. It is useful to separate these areas automatically for several practical applications. For example, in Web data mining, e.g., classification and clustering, identifying main content areas or removing noisy blocks (e.g., advertisements, navigation panels, etc) enables one to produce much better results. It was shown in [51, 52] that the information contained in noisy blocks can seriously harm Web data mining. Another application is Web browsing using a small screen device, such as a PDA. Identifying different content blocks allows one to re-arrange the layout of the page so that the main contents can be seen easily without losing any other information from the page. In this past two years, several papers have been published on this topic [51, 52, 53, 47, 44]. This research also includes detecting common layout and templates of Web pages [4, 25, 45].

2.6 Mining Web Opinion Sources

Consumer opinions used to be very difficult to obtain before the Web was available. Companies usually conduct consumer surveys or engage external consultants to find such opinions about their products and those of their competitors. Now much of the information is publicly available on the Web. There are numerous Web sites and pages containing consumer opinions, e.g., customer reviews of products, forums, discussion groups, and blogs. This online word-of-mouth behavior represents new and measurable sources of information for marketing intelligence. Techniques are now being developed to exploit these sources to help companies and individuals to gain such information effectively and easily [1, 16, 24, 38, 40, 41, 48, 49]. For instance, [24] proposes a feature based summarization method to automatically analyze consumer opinions in customer reviews from online merchant sites and dedicated review sites. The result of such a summary is useful to both potential customers and product manufacturers.

3. PAPERS IN THIS SPECIAL ISSUE

This special issue of SIGKDD Explorations brings together some of the latest research results in the field of Web content mining. It presents eight papers, which deal with a wide range of problems.

All the papers propose some novel and/or principled techniques to solve these problems.

The first paper by Zhang, Lakshmanan, and Zamar studies the problem of extracting data records from a large set of Web pages. What is interesting about this paper is that it proposes a novel method to estimate the current coverage of the results of the system, based on capture-recapture models with unequal capture probabilities. Techniques are also proposed to estimate the error rate of the extracted information. To evaluate the method and ideas proposed in this paper, a large number of experiments have been conducted to demonstrate their effectiveness.

The second paper by Song et al proposes a new method to segment a Web page into blocks of different importance using machine learning methods. Specifically, it proposes a set of features for learning a block importance model for Web pages. As mentioned earlier, page segmentation is important because in many applications, only the main content blocks are useful. Other noisy blocks should be removed. Earlier work in this area requires multiple pages from a site to detect templates [52]. The technique in this paper works on a single page, which is a main advantage.

The third paper by Cimiano and Staab proposes a methodology for extracting knowledge from the Web in knowledge acquisition. Specifically, it reports a system, called PANKOW, which classifies concepts into a given ontology. The method first generates a set of language patterns from the query concept and then uses a search engine (it uses Google) to collect statistical information about each pattern on the Web. The knowledge engineer then makes the decision regarding the classification. The basic idea is that certain lexico-syntactic patterns matched in texts convey a specific semantic relation.

The fourth paper by Zhang et al explores the problem of correlated summarization of a pair of online news articles. The algorithm aligns (sub)topics of the two news articles and summarizes their correlation by sentence extraction. They model a pair of news articles with a weighted bipartite graph. A mutual reinforcement principle is then applied to identify a dense sub-graph of the weighted bipartite graph. Sentences corresponding to the sub-graph are correlated well in textual contents and convey the dominant shared topic of the articles. Their experiment results show that the technique works well.

The fifth paper by Gruhl et al studies the diffusion or the dynamics of information in the blogspace. They show how by using macro (topical) and micro (individual) models, various structures and behaviors can be understood, ranging from the strong driving effect of outside world events on what is being discussed to the applicability of traditional sociological models of influence to bloggers. The discovered characterizations of information propagation allow practical applications to take advantage of these emerging Web phenomena.

The sixth paper by Dong, Madhavan and Halevy presents techniques for searching and matching Web services by exploiting statistics in a large corpus of structures (e.g., WSDL files). Web services are loosely coupled software components, published, located, and invoked across the Web with SOAP. A Web service typically comprises of several operations with parameters. This paper reviews and compares two recent works: Searching for Web services and schema matching, both of which leverage corpora of structures to bridge semantic heterogeneity.

The seventh paper by Sarawagi and Vydiswaran addresses the problem of finding the paths leading to specific "goal pages" on a large Web site. It addresses this problem as sequential labeling with Conditional Random Fields. Thus, unlike prior "focused crawling" works, this paper proposes to capture the dependency or correlation between classifications of sequential pages on a path. The specific focus of "crawling in a Web site" also distinguishes the work from general-purpose focused crawling.

The eighth paper by Chang, He, and Zhang studies dynamic "on-the-fly" semantics discovery for large scale integration on the "deep Web," and proposes "holistic mining" as a conceptual framework unifying initial works and a general approach for such large-scale integration. It reports three sample tasks as evidences: interface extraction, schema matching, and query translation. To generalize, it then proposes holistic mining as a unified insight to observe and leverage "hidden regularities" across holistic sources for large scale integration, and outlines future challenges.

4. SUMMARY

In summary, the eight papers represent some of the latest and most promising research results in this new and exciting field, which continues to make significant impact on real-world applications. We are confident that this special issue will stimulate further research in this area.

5. REFERENCES

[1] Agrawal, R., Rajagopalan, S., Srikant, R., Xu, Y. Mining Newsgroups Using Networks Arising from Social Behavior. *WWW-03*, 2003.

[2] Agrawal, R., Srikant, R. On Integrating Catalogs. *WWW-01*, 2001.

[3] Arasu, A. and Garcia-Molina, H. Extracting Structured Data from Web Pages. *SIGMOD-03*, 2003.

[4] Bar-Yossef, Z. and Rajagopalan, S. Template Detection via Data Mining and its Applications, *WWW-02*, 2002.

[5] Bergman, M. K. The Deep Web: Surfacing Hidden Value. Technical report, BrightPlanet LLC, Dec. 2000

[6] Bunescu, R., Mooney, R. Collective Information Extraction with Relational Markov Networks. *ACL-2004*, 2004.

[7] Califf, M and Mooney, R. Bottom-up Relational Learning of Pattern Matching Rules for Information Extraction. *Journal of Machine Learning Research*, 4:177–210, 2003.

[8] Chakrabarti, S . *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann Publishers, 2002.

[9] Chang, C-H., Lui, S-L. IEPAD: Information Extraction Based on Pattern Discovery. *WWW-10*, 2001.

[10] Chang, K. C.-C., He, B., Li, C., Patel, M., Zhang, Z. Structured Databases on the Web: Observations and Implications. *SIGMOD Record*, 33(3), 2004

[11] Chuang, S.-L. and Chien, L.-F., A Practical Web-based Approach to Generating Topic Hierarchy for Text Segments. *CIKM-04*, 2004.

[12] Cimiano, P., Handschuh, S., Staab, S. Towards the Self-Annotating Web. *WWW-04*, 2004.

[13] Cohen, W., Hurst, M., and Jensen, L. A Flexible Learning System for Wrapping Tables and Lists in HTML Documents.

WWW-02, 2002.

[14] Cohen, W., Sarawagi, S. Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods. *KDD-04*, 2004.

[15] Crescenzi, V., Mecca, G. and Merialdo, P. ROADRUNNER: Towards Automatic Data Extraction from Large Web Sites. *VLDB-01*, 2001.

[16] Dave, K., Lawrence, S., and Pennock, D. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *WWW'03*, 2003.

[17] Doan, A. Madhavan, J. Domingos, P., and Halevy, A. Learning to Map between Ontologies on the Semantic Web. *WWW-02*, 2002.

[18] Freitag, D and McCallum, A. Information Extraction with HMM Structures Learned by Stochastic Optimization. *AAAI-00*, 2000.

[19] Embley, D., Jiang, Y and Ng, Y. Record-boundary discovery in Web documents. *SIGMOD-99*, 1999.

[20] Etzioni, O, Cafarella, M, Downey, D., Kok, S. Popescu, A. Shaked, T., Soderland, S. Weld, S. Web-Scale Information Extraction in KnowItAll (Preliminary Results). *WWW-2004*.

[21] He, B., Chang, K. C.-C. Statistical Schema Matching across Web Query Interfaces. *SIGMOD-03*, 2003.

[22] He, B., Chang, K. C.-C., Han J. Discovering Complex Matchings across Web Query Interfaces: A Correlation Mining Approach. *KDD-04*, 2004.

[23] He, H, Meng, W., Yu, C. Wu, Z. WISE-Integrator: An Automatic Integrator of Web Search Interfaces for E-Commerce. *VLDB-03*, 2003.

[24] Hu, M and Liu, B. Mining and Summarizing Customer Reviews. *KDD-04*, 2004.

[25] Kao, J., Lin, S. Ho, J. Chen, M. Entropy-based Link Analysis for Mining web Informative Structures. *CIKM 2002*.

[26] Kummamuru, K., Lotlikar, R., Roy, S., Singal, K. and Krishnapuram, R., A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. *WWW-04*, 2004.

[27] Kushmerick, N. Wrapper Induction: Efficiency and Expressiveness. *Artificial Intelligence*, 118:15-68, 2000.

[28] Kwok, C., Etzioni, O., Weld, D. Scaling Question Answering to the Web. *WWW-00*, 2000.

[29] Lafferty, J., McCallum, A. Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling or Sequence Data. *ICML-01*, 2001.

[30] Lawrie, D.J. and Croft, W.B., Generating Hierarchical Summaries for Web Searches. *WWW'03*, 2003.

[31] Lerman, K., Getoor L., Minton, S. and Knoblock, C. Using the Structure of Web Sites for Automatic Segmentation of Tables. *SIGMOD-04*, 2004.

[32] Liu, B., Chin, C., Ng, H-T. Mining Topic-Specific Concepts and Definitions on the Web. *WWW-03*, 2003.

[33] Liu, B., Grossman, R. and Zhai, Y. Mining Data Records in Web Pages. *KDD-03*, 2003.

[34] Liu, B., Zhao, K., and Yi, L. Visualizing Web site Comparisons. *WWW-02*, 2002.

[35] Liu, B. Ma, Y., Yu, P. Discovering Unexpected Information from Your Competitors' Web Sites. *KDD-01*, 2001.

[36] Maedche, A. and Staab, A. Mining Ontologies from Text. *12th International Conference on Knowledge Engineering and Knowledge Management*, 2000.

[37] McCallum, A and Li, W. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. *CoNLL-03*, 2003.

[38] Morinaga, S., Yamanishi, K., Tateishi, K, and Fukushima, T. Mining Product Reputations on the Web. *KDD-02*, 2002.

[39] Muslea, I., Minton, S. and Knoblock, C. A Hierarchical Approach to Wrapper Induction. *Agents-99*, 1999.

[40] Nigam, K. and Hurst, M. Towards a Robust Metric of Opinion. *AAAI Spring Symposium on Exploring Attitude and Affect in Text*. 2004.

[41] Pang, B., Lee, L., and Vaithyanathan, S., Thumbs up? Sentiment Classification Using Machine Learning Techniques. *EMNLP-02, 2002.*

[42] Pinto, D., McCallum, A., Wei, X. and Bruce, W. Table Extraction Using Conditional Random Fields. *SIGIR-03*.

[43] Radev, D., Fan, W., Qi, H., Wu, H., Grewal, A. Probabilistic Question Answering on the Web. *WWW-02*, 2002.

[44] Ramaswamy, L, Ivengar, A, Liu, L, and Douglis, F. Automatic Detection of Fragments in Dynamically Generated Web pages. *WWW-04*, 2004.

[45] Reis, D. Golgher, P, Silva, A. Laender, A. Automatic Web News Extraction Using Tree Edit Distance, *WWW-04*, 2004.

[46] Sarawagi, S., Cohen, W. Semi-Markov Conditional Random Fields for Information Extraction, *NIPS-04*, 2004.

[47] Song, R., Liu, H., Wen, J.-R, W.-Y, Ma. Learning Block Importance Models for Web Pages. *WWW-04*, 2004.

[48] Turney, P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *ACL-02*, 2002.

[49] Wilson, T, Wiebe, J, & Hwa, R. Just How Mad are You? Finding Strong and Weak Opinion Clauses. *AAAI-04,* 2004.

[50] Wu, W., Yu, C., Doan, A and Meng, W. An Interactive Clustering-based Approach to Integrating Source Query Interfaces on the Deep Web. *SIGMOD-04*, 2004.

[51] Yi, L. and Liu, B. Eliminating Noisy Information in Web Pages for Data Mining. *KDD-03*, 2003.

[52] Yi, L., and Liu, B. Web Page Cleaning for Web Mining through Feature Weighting *IJCAI-03*, 2003.

[53] Yin, X. and Lee, W. S., Using Link Analysis to Improve Layout on Mobile Devices*. WWW-04, 2004.*

[54] Yu, S., Cai, D., Wen, J.-R. and Ma, W.-Y., Improving Pseudo-Relevance Feedback in Web Information Retrieval Using Web Page Segmentation. *WWW-03*, 2003.

[55] Zamir, O. and Etzioni, O., Grouper: A Dynamic Clustering Interface to Web Search Results. *WWW8*, 1999.

[56] Zeng, H. He, Q, Chen, Z., Ma, W. and Ma, J. Learning to Cluster Web Search Results. *SIGIR-04*, 2004.

[57] Zhang, Z., He, B., Chang, K. C.-C. Understanding Web Query Interfaces: Best-Effort Parsing with Hidden Syntax. *SIGMOD-04*, 2004.