HANDBOOK OF RESEARCH ON

TEXT AND WEB MINING TECHNOLOGIES

Min Song & Yi-fang Brook Wu

Volume I

Detailed Table of Contents

Document keyphrases provide semantic metadata which can characterize documents and produce an overview of the content of a document. This chapter describes a keyphrase identification program (KIP), which extracts document keyphrases by using prior positive samples of human identified domain keyphrases to assign weights to the candidate keyphrases. The logic of our algorithm is: the more keywords a candidate keyphrase contains and the more significant these keywords are, the more likely this candidate phrase is a keyphrase. To obtain human identified positive inputs, KIP first populates its glossary database using manually identified keyphrases and keywords. It then checks the composition of all noun phrases extracted from a document, looks up the database and calculates scores for all these noun phrases. The ones having higher scores will be extracted as keyphrases. KIP's learning function can enrich the glossary database by automatically adding new identified keyphrases to the database.

This chapter proposes a text mining model to handle shallow text representation and processing for mining purposes in an integrated way. Its aims are to look for interesting explanatory knowledge across text documents. The proposed model involves a mixture of different techniques from evolutionary computation and other kinds of text mining methods.

Section II
Classification and Clustering

Syllabi are important educational resources. Gathering syllabi that are freely available and creating useful services on top of the collection presents great value for the educational community. However, searching for a syllabus on the Web using a generic search engine is an error-prone process and often yields too many irrelevant links. In this chapter, we describe our empirical study on automatic syllabus classification using dupport vector machines (SVM) to filter noise out from search results. We describe various steps in the classification process from training data preparation, feature selection, and classifier building using SVMs. Empirical results are provided and discussed. We hope our reported work will also benefit people who are interested in building other genre-specific repositories.

In traditional text categorization, a classifier is built using labeled training documents from a set of predefined classes. This chapter studies a different problem: partially supervised text categorization. Given a set P of positive documents of a particular class and a set U of unlabeled documents (which contains both hidden positive and hidden negative documents), we build a classifier using P and U to classify the data in U as well as future test data. The key feature of this problem is that there is no labeled negative document, which makes traditional text classification techniques inapplicable. In this chapter, we introduce the main techniques S-EM, PEBL, Roc-SVM and A-EM, to solve the partially supervised problem. In many application domains, partially supervised text categorization is preferred since it saves on the labor-intensive effort of manual labeling of negative documents.

Chapter VI

Mining techniques can play an important role in automatic image classification and content-based retrieval. A novel method for image classification based on feature element through association rule mining is presented in this chapter. The effectiveness of this method comes from two sides. The visual meanings of images can be well captured by discrete feature elements. The associations between the description features and the image contents can be properly discovered with mining technology. Experiments with real images show that the new approach provides not only lower classification and retrieval error but also higher computation efficiency.

Chapter VII

This chapter introduces two practical techniques for improving Naïve Bayes text classifiers that are widely used for text classification. The Naïve Bayes has been evaluated to be a practical text classification algorithm due to its simple classification model, reasonable classification accuracy, and easy update of classification model. Thus, many researchers have a strong incentive to improve the Naïve Bayes by combining it with other meta-learning approaches such as EM (Expectation Maximization) and Boosting. The EM approach is to combine the Naïve Bayes with the EM algorithm and the Boosting approach is to use the Naïve Bayes as a base classifier in the AdaBoost algorithm. For both approaches, a special uncertainty measure fit for Naïve Bayes learning is used. In the Naïve Bayes learning framework, these approaches are expected to be practical solutions to the problem of lack of training documents in text classification systems.

Chapter VIII

In this chapter, we are interested in proteins classification starting from their primary structures. The goal is to automatically affect proteins sequences to their families. The main originality of the approach is that we directly apply the text categorization framework for the protein classification with very minor

modifications. The main steps of the task are clearly identified: we must extract features from the unstructured dataset, we use the fixed length n-grams descriptors; we select and combine the most relevant one for the learning phase; and then, we select the most promising learning algorithm in order to produce accurate predictive model. We obtain essentially two main results. First, the approach is credible, giving accurate results with only 2-grams descriptors length. Second, in our context where many irrelevant descriptors are automatically generated, we must combine aggressive feature selection algorithms and low variance classifiers such as SVM (Support Vector Machine).

Chapter IX

Feature-based semantic measurements have played a dominant role in conventional data clustering algorithms for many existing applications. However, the applicability of existing data clustering approaches to wider range of applications is limited due to issues such as complexity involved in semantic computation, long pre-processing time required for feature preparation, and poor extensibility of semantic measurement due to non-incremental feature source. This chapter first summarises the many commonly used clustering algorithms and feature-based semantic measurements, and then highlights the shortcomings to make way for the proposal of an adaptive clustering approach based on featureless semantic measurements. The chapter concludes with experiments demonstrating the performance and wide applicability of the proposed clustering approach.

Chapter X

In this chapter, we introduce three nature inspired swarm intelligence clustering approaches for document clustering analysis. The major challenge of today's information society is being overwhelmed with information on any topic they are searching for. Fast and high-quality document clustering algorithms play an important role in helping users to effectively navigate, summarize, and organize the overwhelmed information. The swarm intelligence clustering algorithms use stochastic and heuristic principles discovered from observing bird flocks, fish schools, and ant food forage. Compared to the traditional clustering algorithms, the swarm algorithms are usually flexible, robust, decentralized, and self-organized. These characters make the swarm algorithms suitable for solving complex problems, such as document clustering.

Chapter XI

Clustering is a process of finding natural grouping present in a dataset. Various clustering methods are proposed to work with various types of data. The quality of the solution as well as the time taken to derive the solution is important when dealing with large datasets like that in a typical documents database. Recently hybrid and ensemble based clustering methods are shown to yield better results than conventional methods. The chapter proposes two clustering methods; one is based on a hybrid scheme and the other based on an ensemble scheme. Both of these are experimentally verified and are shown to yield better and faster results.

Chapter XII

The classification of textual documents has been the subject of many studies. Technologies like the Web and numerical libraries facilitated the exponential growth of available documentation. The classification of textual documents is very important since it allows the users to effectively and quickly fly over and understand better the contents of large corpora. Most classification approaches use the supervised method of training, more suitable with small corpora and when human experts are available to generate the best classes of data for the training phase, which is not always feasible. The unsupervised classification or "clustering" methods make emerge latent (hidden) classes automatically with minimum human intervention, There are many, and the self organized maps (SOM) by Kohonen is one of the algorithms for unsupervised classification that gather a certain number of similar objects in groups without a priori knowledge. This chapter introduces the concept of unsupervised classification of textual documents and proposes an experiment with a conceptual approach for the representation of texts and the method of Kohonen for clustering.

Chapter XIII

This chapter proposes a Web mining system based on back-propagation neural network to support users for decision making. To handle the scalability issue of the Web mining system, the proposed system provides a multi-agent based neural network system in a parallel way.

Section III
Database, Ontology, and the Web

With the emergence of XML standardization, XML documents have been widely used and accepted in almost all the major industries. As a result of the widespread usage, it has been considered essential to not only store these XML documents but also to mine them to discover useful information from them. One of the very popular techniques to mine XML documents is frequent pattern mining, which has huge potential in varied domains such as bio-informatics, network analysis. This chapter presents some of the existing techniques to discover frequent patterns from XML documents. It also covers the applications and addresses the major issues in mining XML documents.

XML has gained popularity for information representation, exchange and retrieval. As XML material becomes more abundant, its heterogeneity and structural irregularity limit the knowledge that can be gained.. The utilisation of data mining techniques becomes essential for improvement in XML document handling. This chapter presents the capabilities and benefits of data mining techniques in the XML domain, as well as, a conceptualization of the XML mining process. It also discusses the techniques that can be applied to XML document structure and/or content for knowledge discovery.

In the context of knowledge discovery in databases, *data reduction* is a pre-processing step delivering succinct yet meaningful data to sequent stages. If the target of mining are data streams, then it is crucial to suitably reduce them, since often analyses on such data require multiple scans. In this chapter, we propose a histogram-based approach to reducing sliding windows supporting approximate arbitrary (i.e., non biased) range-sum queries. The histogram is based on a hierarchical structure (as opposed to the flat structure of traditional ones) and it results suitable to directly support hierarchical queries, such as drill-down and roll-up operations. In particular, both sliding window shifting and quick query answering operations are logarithmic in the sliding window size. Experimental analysis shows the superiority of our method in terms of accuracy w.r.t. the state-of-the-art approaches in the context of histogram-based sliding window reduction techniques.

tion of information from these two kinds of sources is crucial for managing and extracting knowledge. By structuring and defining the concepts and relationships within a biomedical domain, BioOntologies have taken a key role in this integration. This chapter describes the role of BioOntologies in sharing, integrating and mining biological information, discusses some of the most relevant BioOntologies and illustrates how they are being used by automatic tools to improve our understanding of life.

Chapter XX

In this chapter, we present a thesaurus application in the field of text mining and more specifically automatic indexing on the set of descriptors defined by a thesaurus. We begin by presenting various definitions and a mathematical thesaurus model, and also describe various examples of real world thesauri which are used in official institutions. We then explore the problem of thesaurus-based automatic indexing by describing its difficulties and distinguishing features and reviewing previous work in this area. Finally, we propose various lines of future research.

Chapter XXI

The goal of this chapter is to present an approach to mine texts through the analysis of higher level characteristics (called "concepts'), minimizing the vocabulary problem and the effort necessary to extract useful information. Instead of applying text mining techniques on terms or keywords labeling or extracted from texts, the discovery process works over concepts extracted from texts. Concepts represent real world attributes (events, objects, feelings, actions, etc.) and, as seen in discourse analysis, they help to understand ideas and ideologies present in texts. A previous classification task is necessary to identify concepts inside the texts. After that, mining techniques are applied over the concepts discovered. The chapter will discuss different concept-based text mining techniques and present results from different applications.

Chapter XXII

This chapter aims at providing an overview about the use of statistical methods supporting the Web usage mining. Within the first part is described the framework of the Web usage mining as a branch of the

Web mining committed to the study of how to use a Website. Then, the data (object of the analysis) are detailed together with the problems linked to the pre-processing. Once clarified, the data origin and their treatment for a correct development of a Web Usage analysis,the focus shifts on the statistical techniques that can be applied to the analysis background, with reference to binary segmentation methods. Those latter allow the discrimination through a response variable that determines the affiliation of the users to a group by considering some characteristics detected on the same users.

Chapter XXIII

This chapter presents a new approach of mining the Web to identify people of similar background. To find similar people from the Web for a given person, two major research issues are person representation and matching persons. In this chapter, a person representation method which uses a person's personal Website to represent this person's background is proposed. Based on this person representation method, the main proposed algorithm integrates textual content and hyperlink information of all the Web pages belonging to a personal Website to represent a person and match persons. Other algorithms are also explored and compared to the main proposed algorithm. The evaluation methods and experimental results are presented.

Chapter XXIV

This chapter describes how Web usage patterns can be used to improve the navigational structure of a Website. The discussion begins with an illustration of visualization tools that study aggregate and individual link traversals. The use of data mining techniques such as classification, association, and sequence analysis to discover knowledge about Web usage, such as navigational patterns, is also discussed. Finally, a graph theoretic algorithm to create an optimal navigational hyperlink structure, based on known navigation patterns, is presented. The discussion is supported by analysis of real-world datasets.

Chapter XXV

This chapter surveys the usage of a modeling language, WebML, for the design and the management of dynamic Web applications. The chapter also reports a case study of the effectiveness of WebML and its conceptual modeling methods by analyzing Web logs. To analyze Web logs, the chapter utilizes the data mining paradigm of item sets and frequent patterns.

Volume II

Automatic reference tracking involves systematic tracking of reference articles listed for a particular research paper by extracting the references of the input seed publication and further analyzing the relevance of the referred paper with respect to the seed paper. This tracking continues recursively with every reference paper being assumed as seed paper at every track level until the system finds any irrelevant (or far relevant) references deep within the reference tracks which does not help much in the understanding of the input seed research paper at hand. The relevance is analysed based on the keywords collected from the title and abstract of the referred article. The objective of the reference tracking system is to automatically list down closely relevant reference articles to aid the understanding of the seed paper thereby facilitating the literature survey of the aspiring researcher. This chapter proposes the system design and evaluation of automatic reference tracking system discussing the observations obtained.

As more electronic text is readily available, and more applications become knowledge intensive and ontology-enabled, term extraction, also known as automatic term recognition or terminology mining is increasingly in demand. This chapter first presents a comprehensive review of the existing techniques, discusses several issues and open problems that prevent such techniques from being practical in real-life applications, and then proposes solutions to address these issues. Keeping afresh with the recent advances in related areas such as text mining, we propose new measures for the determination of unithood, and a new scoring and ranking scheme for measuring termhood to recognise domain-specific terms. The chapter concludes with experiments to demonstrate the advantages of our new approach.

Over 60% of the online population are non-English speakers and it is probable the number of non-English speakers is growing faster than English speakers. Most search engines were originally engineered for English. They do not take full account of inflectional semantics nor, for example, diacritics or the use of capitals. The main conclusion from the literature is that searching using non-English and non-Latin based queries results in lower success and requires additional user effort so as to achieve acceptable recall and precision. In this chapter a Greek query log is morphologically and grammatically analyzed and a number of queries are submitted to search engines and their relevance is evaluated with the aid of real users. A Greek meta-searcher redirecting normalized queries to Google.gr is also presented and evaluated. An increase in relevance is reported when stopwords are eliminated and queries are normalized based on their morphology.

Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI), when applied to information retrieval, has been a major analysis approach in text mining. It is an extension of the vector space method in information retrieval, representing documents as numerical vectors but using a more sophisticated mathematical approach to characterize the essential features of the documents and reduce the number of features in the search space. This chapter summarizes several major approaches to this dimensionality reduction, each of which has strengths and weaknesses, and it describes recent breakthroughs and advances. It shows how the constructs and products of LSA applications can be made user-interpretable and reviews applications of LSA beyond information retrieval, in particular, to text information visualization.

Text mining systems such as categorizers and query retrievers of the first generation were largely hinged on word level statistics and provided a wonderful first-cut approach. However systems based on simple word-level statistics quickly saturate in performance, despite the best data mining and machine learning algorithms. This problem can be traced to the fact that, typically, naive, word-based feature representations are used in text applications, which prove insufficient in bridging two types of chasms within and across documents, viz. lexical chasm and syntactic chasm . The latest wave in text mining technology has been marked by research that will make extraction of subtleties from the underlying meaning of text, a possibility. In the following two chapters, we pose the problem of underlying meaning extraction from text documents, coupled with world knowledge, as a problem of bridging the chasms by exploiting associations between entities. The entities are words or word collocations from documents. We utilize two types of entity associations, viz. paradigmatic (PA) and syntagmatic (SA). We present first-tier algorithms that use these two word associations in bridging the syntactic and lexical chasms. We also propose second-tier algorithms in two sample applications, viz., question answering and text classification which use the first-tier algorithms. Our contribution lies in the specific methods we introduce for exploiting entity association information present in WordNet, dictionaries, corpora and parse trees for improved performance in text mining applications.

This chapter surveys well-known Web content mining techniques that can be used for addressing the problem of providing personalized access to the contents of Usenet community. It also discusses how the end-results of knowledge discovery process from the Usenet sites are utilized by individual users.

Section V
Application and Survey

Searching for frequent pieces in a database with some sort of text is a well-known problem. A special sort of text is program code as e.g. C++ or machine code for embedded systems. Filtering out duplicates in large software projects leads to more understandable programs and helps avoiding mistakes when reengineering the program. On embedded systems the size of the machine code is an important issue. To ensure small programs, duplicates must be avoided. Several different approaches for finding code duplicates based on the text representation of the code or on graphs representing the data and control flow of the program and graph mining algorithms.

In Golbeck and Hendler (2006), authors consider those social friendship networking sites where users explicitly provide trust ratings to other members. However, for large social friendship networks it is infeasible to assign trust ratings to each and every member so they propose an inferring mechanism which would assign binary trust ratings (trustworthy/non-trustworthy) to those who have not been assigned one. They demonstrate the use of these trust values in e-mail filtering application domain and report encouraging results. Authors also assume three crucial properties of trust for their approach to work: transitivity, asymmetry, and personalization. These trust scores are often transitive, meaning, if

Alice trusts Bob and Bob trusts Charles then Alice can trust Charles. Asymmetry says that for two people involved in a relationship, trust is not necessarily identical in both directions. This is contrary to what was proposed in Yu and Singh (2003). They assume symmetric trust values in the social friendship network. Social networks allow us to share experiences, thoughts, opinions, and ideas. Members of these networks, in return experience a sense of community, a feeling of belonging, a bonding that members matter to one another and their needs will be met through being together. Individuals expand their social networks, convene groups of like-minded individuals and nurture discussions. In recent years, computers and the World Wide Web technologies have pushed social networks to a whole new level. It has made possible for individuals to connect with each other beyond geographical barriers in a "flat" world. The widespread awareness and pervasive usability of the social networks can be partially attributed to Web 2.0. Representative interaction Web services of social networks are social friendship networks, the blogosphere, social and collaborative annotation (aka "folksonomies"), and media sharing. In this work, we briefly introduce each of these with focus on social friendship networks and the blogosphere. We analyze and compare their varied characteristics, research issues, state-of-the-art approaches, and challenges these social networking services have posed in community formation, evolution and dynamics, emerging reputable experts and influential members of the community, information diffusion in social networks, community clustering into meaningful groups, collaboration recommendation, mining "collective wisdom" or "open source intelligence" from the exorbitantly available user-generated contents. We present a comparative study and put forth subtle yet essential differences of research in friendship networks and Blogosphere, and shed light on their potential research directions and on cross-pollination of the two fertile domains of ever expanding social networks on the Web.

Chapter XXXVII

In this chapter we present an information system conceived for supporting managers of Public Health Care Agencies to decide the new health care services to propose. Our system is HL7-aware; in fact, it uses the HL7 (Health Level Seven) standard (Health Level Seven [HL7], 2007) to effectively handle the interoperability among different Public Health Care Agencies. HL7 provides several functionalities for the exchange, the management and the integration of data concerning both patients and health care services. Our system appears particularly suited for supporting a rigorous and scientific decision making activity, taking a large variety of factors and a great amount of heterogeneous information into account.

Chapter XXXVIII

Building effective multi-target classifiers is still an on-going research issue: this chapter proposes the use of the knowledge gleaned from a human expert as a practical way for decomposing and extend the

proposed binary strategy. The core is a greedy feature selection approach that can be used in conjunction with different classification algorithms, leading to a feature selection process working independently from any classifier that could then be used. The procedure takes advantage from the Minimum Description Length principle for selecting features and promoting accuracy of multi-target classifiers. Its effectiveness is asserted by experiments, with different state-of-the-art classification algorithms such as Bayesian and Support Vector Machine classifiers, over dataset publicly available on the Web: gene expression data from DNA micro-arrays are selected as a paradigmatic example, containing a lot of redundant features due to the large number of monitored genes and the small cardinality of samples. Therefore, in analysing these data, like in text mining, a major challenge is the definition of a feature selection procedure that highlights the most relevant genes in order to improve automatic diagnostic classification.

Chapter XXXIX

Text mining is an instrumental technology that today's organizations can employ to extract information and further evolve and create valuable knowledge for more effective knowledge management. It is also an important tool in the arena of information systems security (ISS). While a plethora of text mining research has been conducted in search of revamped technological developments, relatively limited attention has been paid to the applicable insights of text mining in ISS. In this chapter, we address a variety of technological applications of text mining in security issues. The techniques are categorized according to the types of knowledge to be discovered and the text formats to be analyzed. Privacy issues of text mining as well as future trends are also discussed.

Chapter XL

This chapter introduces collaborative filtering-based recommendation systems, which has become an integral part of e-commerce applications, as can be observed in sites like Amazon.com. It will present several techniques that are reported in the literature to make useful recommendations, and study their limitations. The chapter also lists the issues that are currently open and the future directions that may be explored to address those issues. Furthermore, the authors hope that understanding of these limitations and issues will help build recommendation systems that are of high accuracy and have few false positive errors (which are products that are recommended, though the user does not like them).

The purpose of this chapter is to provide an overview of prevalent measures for evaluating the quality of system output in seven key text mining task domains. For each task domain, a selection of widely used, well applicable measures is presented, and their strengths and weaknesses are discussed. Performance evaluation is essential for text mining system development and comparison, but the selection of a suitable performance evaluation measure is not a straightforward task. Therefore this chapter also attempts to give guidelines for measure selection. As measures are under constant development in many task domains and it is important to take the task domain characteristics and conventions into account, references to relevant performance evaluation events and literature are provided.

The biology literatures have been increased in an exponential growth in recent year. The researchers need an effective tool to help them find out the needed information in the databases. Text mining is a powerful tool to solve this problem. In this chapter, we talked about the features of text mining and bioinformatics, text mining applications, research methods in bioinformatics and problems and future path.

With the increased use of Internet, a large number of consumers first consult on line resources for their healthcare decisions. The problem of the existing information structure primarily lies in the fact that the vocabulary used in consumer queries is intrinsically different from the vocabulary represented in medical literature. Consequently, the medical information retrieval often provides poor search results. Since consumers make medical decisions based on the search results, building an effective information retrieval system becomes an essential issue. By reviewing the foundational concepts and application components of medical information retrieval, this chapter will contribute to a body of research that seeks appropriate answers to a question like "How can we design a medical information retrieval system that can satisfy consumer's information needs?"

Chapter XLIV

This chapter presents background on text mining, and comparisons and summaries of seven selected software for text mining. The text mining software selected for discussion and comparison in this chapter are: Compare Suite by AKS-Labs, SAS Text Miner, Megaputer Text Analyst, Visual Text by Text Analysis International, Inc. (TextAI), Magaputer PolyAnalyst, WordStat by Provalis Research, and SPSS Clementine. This chapter not only discusses unique features of these text mining software packages but also compares the features offered by each in the following key steps in analyzing unstructured qualitative data: data preparation, data analysis, and result reporting. A brief discussion of Web mining and its software are also presented, as well as conclusions and future trends.

Chapter XLV

This chapter describes the application of several text mining techniques to discover patterns in the health insurance schedule with an aim to uncover any inconsistency or ambiguity in the schedule. Based on the survey, this chapter experiments with classification and clustering on full features and reduced features using the latent semantic kernel algorithm. The results show that the LSK algorithm works well on Health Insurance Commission schedules.

Chapter XLVI

This chapter proposes a Web mining system that incorporates both online efficiency and off-line effectiveness to provide the right information based on users' preferences. The proposed system is applied to the Web site marketing of mobile phones. Through this case study, this chapter demonstrates that a

query-response containing a reasonable number of mobile phones best matched a user's preferences can be provided.

Chapter XLVII

Text mining technology can be used to assist in finding relevant or novel information in large volumes of unstructured data, such as that which is increasingly available in the electronic scientific literature. However, publishers are not text mining specialists, nor typically are the end-user scientists who consume their products. This situation suggests a Web services based solution, where text mining specialists process the literature obtained from publishers and make their results available to remote consumers (research scientists). In this chapter we discuss the integration of Web services and text mining within the domain of scientific publishing and explore the strengths and weaknesses of three generic architectural designs for delivering text mining Web services. We argue for the superiority of one of these and demonstrate its viability by reference to an application designed to provide access to the results of text mining over the PubMed database of scientific abstracts.