1

Overview

Anne Kao and Stephen R. Poteet

1.1 Introduction

Text mining is the discovery and extraction of interesting, non-trivial knowledge from free or unstructured text. This encompasses everything from information retrieval (i.e., document or web site retrieval) to text classification and clustering, to (somewhat more recently) entity, relation, and event extraction. Natural language processing (NLP), is the attempt to extract a fuller meaning representation from free text. This can be put roughly as figuring out who did what to whom, when, where, how and why. NLP typically makes use of linguistic concepts such as part-of-speech (noun, verb, adjective, etc.) and grammatical structure (either represented as phrases like noun phrase or prepositional phrase, or dependency relations like subject-of or object-of). It has to deal with anaphora (what previous noun does a pronoun or other back-referring phrase correspond to) and ambiguities (both of words and of grammatical structure, such as what is being modified by a given word or prepositional phrase). To do this, it makes use of various knowledge representations, such as a lexicon of words and their meanings and grammatical properties and a set of grammar rules and often other resources such as an ontology of entities and actions, or a thesaurus of synonyms or abbreviations.

This book has several purposes. First, we want to explore the use of NLP techniques in text mining, as well as some other technologies that are novel to the field of text mining. Second, we wish to explore novel ways of integrating various technologies, old or new, to solve a text mining problem. Next, we would like to look at some new applications for text mining. Finally, we have several chapters that provide various supporting techniques for either text mining or NLP or both, or enhancements to existing techniques.

1.2 Approaches that Use NLP Techniques

The papers in our first group deal with approaches that utilize to various degrees more in-depth NLP techniques. All of them use a parser of some sort or another, one of them uses some morphological analysis (or rather generation), and two of them use other lexical resources, such as WordNet, FrameNet, or VerbNet. The first three use off-the-shelf parsers while the last uses their own parser.

Popescu and Etzioni combine a wide array of techniques. Among these are NLP techniques such as parsing with an off-the-shelf parser, MINIPAR, morphological rules to generate nouns from adjectives, and WordNet (for its synonymy and antonymy information, its IS-A hierarchy of word meanings, and for its adjective-to-noun pertain relation). In addition, they use hand-coded rules to extract desired relations from the structures resulting from the parse. They also make extensive and key use of a statistical technique, pointwise mutual information (PMI), to make sure that associations found both in the target data and in supplementary data downloaded from the Web are real. Another distinctive technique of theirs is that they make extensive use of the Web as a source of both word forms and word associations. Finally, they introduce relaxation labeling, a technique from the field of image-processing, to the field of text mining to perform context sensitive classification of words.

Bunescu and Mooney adapt Support Vector Machines (SVMs) to a new role in text mining, namely relation extraction, and in the process compare the use of NLP parsing with non-NLP approaches. SVMs have been used extensively in text mining but always to do text classification, treating a document or piece of text as an unstructured bag of words (i.e., only what words are in the text and what their counts are, not their position with respect to each other or any other structural relationships among them). The process of extracting relations between entities, as noted above, has typically been presumed to require parsing into natural language phrases. This chapter explores two new kernels for SVMs, a subsequence kernel and a dependency path kernel, to classify the relations between two entities (they assume the entities have already been extracted by whatever means). Both of these involve using a wholly novel set of features with an SVM classifier. The dependency path kernel uses information from a dependency parse of the text while the subsequence kernel treats the text as just a string of tokens. They test these two different approaches on two different domains and find that the value of the dependency path kernel (and therefore of NLP parsing) depends on how well one can expect the parser to perform on text from the target domain, which in turn depends on how many unknown words and expressions there are in that domain.

Mustafaraj et al. also combine parsing with statistical approaches to classification. In their case they are using an ensemble or committee of three different classifiers which are typically used with non-NLP features but the features they use are based on parse trees. In addition, their application requires a morphological analysis of the words in their domain, given the nature of German, their target language. They explore the use of off-the-shelf POS taggers and morphological analyzers for this purpose, but find them falling short in their domain (a technical one, electrical fault diagnosis), and have to result to hand coding the morphological rules. Another couple of NLP resources that they utilize are FrameNet and VerbNet to find relevant verbs and relationships to map into their knowledge-engineering categories, but this is used off-line for analysis rather than in on-line processing. Finally, they use active learning to efficiently train their classifiers, a statistical technique that is relatively new to text mining (or data mining in general, for that matter).

Marchisio et al. utilize NLP techniques almost exclusively, writing their own parser to do full parsing and using their novel indexing technique to compress complex parse forests in a way that captures basic dependency relations like subject-of, object-of, and verb-modification like time, location, etc., as well as extended relations involving the modifiers of the entities involved in the basic relations or other entities associated with them in the text or in background knowledge. The index allows them to rapidly access all of these relations, permitting them to be used in document search, an area that has long been considered not to derive any benefit from any but surface NLP techniques like tokenization and stemming. This entails a whole new protocol for search, however, and the focus of their article is on how well users adapt to this new protocol.

1.3 Non-NLP Techniques

Boontham et al. discuss the use of three different approaches to categorizing the free text responses of students to open-ended questions: simple word matching, Latent Semantic Analysis (LSA), and a variation on LSA which they call Topic Models. LSA and Topic Models are both numerical methods for generating new features based on linear algebra and ultimately begin with a representation of the text as a bag of words. In addition, they use discriminant analysis from statistics for classification. Stemming and soundex (a method for correcting misspelling by representing words in a way that roughly corresponds to their pronunciation) are used in the word matching component. Stemming is the only NLP technique used.

McCarthy et al. also use LSA as their primary technique, employing it to compare different sections of a document rather than whole documents and develop a “signature” of documents based on the correlation between different sections.

Schmidtler and Amtrup combine an SVM with a Markov chain to determine how to separate sequences of text pages into distinct documents of different types given that the text pages are very noisy, being the product of optical character recognition. They do a nice job of exploring the different ways they might model a sequence of pages, in terms both of what categories one might assign to pages and how to combine page content and sequence information. They use simple techniques like tokenization and stemming, but not more complex NLP techniques.

Atkinson uses a technique that is very novel for text mining, genetic algorithms (GAs). Genetic algorithms are typically used for solving problems where the features can be represented as binary vectors. Atkinson adapts this to text representations by employing a whole range of numerical and statistical methods, including LSA and Markov chains, and various metrics build on these. However, other than some manually constructed contexts for rhetorical roles, he uses no true NLP techniques.

1.4 Range of Applications

The papers in this book perform a wide range of applications, some more traditional for text mining and some quite novel.

Marchisio et al. take a novel approach to a very traditional application, simple search or document retrieval. They introduce a new paradigm, taking advantage of the linguistic structure of the documents as opposed to key words. Their end-user is the average user of a web search engine.

There are several variants on information extraction.

Bunescu and Mooney look at extracting relations, which, along with entity extraction, is an important current research area in text mining. They focus on two domains, bioinformatics and newspaper articles, each involving a completely different set of entities and relations. The former involves entities like genes, proteins, and cells, and relations like protein-protein interactions and subcellular localization. The latter involves more familiar entities like people, organizations, and locations and relations like "belongs to," "is head of," etc.

Mustafaraj et al. focus on extracting a different kind of relation, the roles of different entities relevant to diagnosis in the technical domain of electrical engineering. These roles include things like "observed object," "symptom," and "cause." In the end, they are trying to mark-up the text of diagnostic reports in a way to facilitate search and the extraction of knowledge about the domain.

Popescu and Etzioni's application is the extraction of product features, parts, and attributes, and customers' or users' opinions about these (both positive and negative, and how strongly they feel) from customer product reviews. These include specialized entities and relations, as well as opinions and their properties, which do not quite fit into these categories.

Atkinson ventures into another novel extraction paradigm, extracting knowledge in form of IF-THEN rules from scientific studies. The scientific domain he focuses on in this particular study is agricultural and food science.

The remaining applications do not fit into any existing text mining niche very well. Schmidtler et al. need to solve a very practical problem, that of separating a stack of pages into distinct documents and labeling the document type. Complicating this problem is the need to use optical character recognition, which results in very noisy text data (lots of errors at the character level). To help overcome this, they utilize whatever sequential information is available in several ways: in setting up the categories (not just document type but beginning/middle/end of document type); in using the category of preceding pages as input in the prediction of a page, and in incorporating knowledge about the number of pages in each document type and hard constraints on the possible sequencing of document types.

McCarthy et al. investigate the use of LSA to compare the similarity of the different sections of scientific studies as a contribution to rhetorical analysis. While the tool is at first blush useful primarily in the scientific field of discourse analysis, they suggest a couple of practical applications, using it to help classify different types of documents (genre and field) or, by authors, to assess how their document measures up to other documents in the same genre and field.

Finally, Boonthum et al. explore the use of various text mining techniques in pedagogy, i.e., to give feedback to students based on discursive rather than categorical (i.e., true-false or multiple choice) answers. In the end, it is a kind of classification problem, but they investigate a method to adapt this quickly to a new domain and set of questions, an essential element for this particular application.

1.5 Supporting Techniques

In addition to various approaches using text mining for some application, there are several papers that explore various techniques that can support text mining (and frequently other data mining) techniques.

Liu et al. investigate a new means of overcoming one of the more important problems in automatic text classification, imbalanced data (the situation where some categories have a lot of examples in the data and other categories have very few examples in the data). They explore various term weighting schemes inspired by the TFIDF metric (term frequency / inverse document frequency) used traditionally in document retrieval in feature selection, and demonstrate that the resulting weighted features show improved performance when used with an SVM.

Brank et al. do a nice survey and classification of approaches to evaluating ontologies for their appropriateness for different domains and tasks, and propose their own metric. Ontologies are an important component for many NLP and text mining applications (e.g., topic classification, entity extraction) and, while the method they propose is based on graph theoretic principles rather than on text mining, many of the other approaches they survey utilize text mining principles as part of the evaluation (or part of the automatic or semi-automatic generation) of ontologies for a particular domain.

Finally, the final chapter by Schmitt et al. is rather different from the other chapters, being more of a tutorial that can benefit students and seasoned professionals alike. It shows how to construct a broad range of text mining and NLP tools using simple UNIX commands and sed and awk (and provides an excellent primer on these in the process). These tools can be used to perform a number of functions, from quite basic ones like tokenization, stemming, or synonym replacement, which are fundamental to many applications, to more complex or specialized ones, like constructing a concordance (a list of terms in context from a corpus, a set of documents to be used for training or analysis) or merging text from different formats to capture important information from each while eliminating irrelevant notations (e.g., eliminating irrelevant formatting mark-up but retaining information relevant both to the pronunciation and kanji forms of different Japanese characters. This information is not only useful for people working on UNIX (or Linux), but can be fairly easily adapted to Perl, which shares much of the regular expression language features and syntax of the UNIX tools, sed and awk.

1.6 Future Work

With the increased use of the Internet, text mining has become increasingly important since the term came into popular usage over 10 years ago. Highly related and specialized fields such as web mining and bioinformatics have also attracted a lot of research work. However, more work is still needed in several major directions. (1) Data mining practitioners largely feel that the majority of data mining work lies in data cleaning and data preparation. This is perhaps even more true in the case of text mining. Much text data does not follow prescriptive spelling, grammar or style rules. For example, the language used in maintenance data, help desk reports, blogs, or email does not resemble that of well-edited news articles at all. More studies on how and to what degree the quality of text data affects different types of text mining algorithms, as well as better methods to 'preprocess' text data would be very beneficial. (2) Practitioners of text mining are rarely sure whether an algorithm demonstrated to be effective on one type of data will work on another set of data. Standard test data sets can help compare different algorithms, but they can never tell us whether an algorithm that performs well on them will perform well on a particular user's dataset. While establishing a fully articulated natural language model for each genre of text data is likely an unreachable goal, it would be extremely useful if researchers could show which types of algorithms and parameter settings tend to work well on which types of text data, based on relatively easily ascertained characteristics of the data (e.g., technical vs. non-technical, edited vs. non-edited, short news vs. long articles, proportion of unknown vs. known words or jargon words vs. general words, complete, well-punctuated sentences vs. a series of phrases with little or no punctuation, etc.) (3) The range of text mining applications is now far broader than just information retrieval, as exhibited by some of the new and interesting applications in this book. Nevertheless, we hope to see an even wider range of applications in the future and to see how they drive additional requirements for text mining theory and methods. In addition, newly emerging fields of study such as link analysis (or link mining) have suggested new directions for text mining research, as well. Our hope is that between new application areas and cross-pollination from other fields, text mining will continue to thrive and see new breakthroughs.