Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 2, February 2015, pg.7 – 13

RESEARCH ARTICLE

Text Data Mining with Different Comparisons

Geetanjali Panda[1], Bhavani Shankar Panda[2], B.Giridhar[3]

[1]M.Tech in CSE, Utkal University, Odisha, India

[2,3]Asst.Prof in CSE Dept, Centurion University, Odisha

[2] chintupanda@gmail.com; [3] giridhar.bcse@gmail.com

Abstract*— Text data mining should be useful for anticipating new technologies and new uses for existing technologies, insofar as one can attempt to connect complementary pieces of information across two different domains, or subsets, of the scientific literature. The possibilities for data mining from large text collections are virtually untapped. Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to decipher automatically. Perhaps for this reason, there has been little work in text data mining to date, and most people who have talked about it have either conflated it with information access or have not made use of text directly to discover heretofore unknown information. In this paper I will first define data mining, information access, and corpus-based computational linguistics, and then discuss the relationship of these to text data mining. The intent behind these contrasts is to draw attention to exciting new kinds of problems for computational linguists. I describe examples of what I consider to be real text data mining efforts and briefly outline our recent ideas about how to pursue exploratory data analysis over text.

I. INTRODUCTION

Technological innovation often proceeds by applying advances made in one field to a separate arena. Once the innovation is implemented, the transfer of knowledge may appear obvious or even inevitable, but without the benefit of hindsight it is surprisingly difficult to identify. Specific technologies that are ripe for transfer[1]. The nascent field of text data mining (TDM) has the peculiar distinction of having a name and a fair amount of hype but as yet almost no practitioners. I suspect this has happened because people assume TDM is a natural extension of the slightly less nascent field of data mining (DM), also known as knowledge discovery in databases, and information archeology. Additionally, there are some disagreements about what actually constitutes data mining. It turns out that ``mining'' is not a very good metaphor for what people in the field actually do. Mining implies extracting precious nuggets of ore from otherwise worthless rock. If data mining really followed this metaphor, it would mean that people were discovering new factoids within their inventory databases. However, in practice this is not really the case. Instead, data mining applications tend to be (semi)automated discovery of trends and patterns across very large datasets[2], usually for the purposes of decision making. Part of what I wish to argue here is that in the case of text, it can be interesting to take the mining-for-nuggets metaphor seriously.

II. TDM VS. INFORMATION ACCESS

This led us to revise the query, to assess whether one or more of these packaging technologies might plausibly. It is important to differentiate between text data mining and information access (or information retrieval, as it is more widely known). The goal of information access is to help users find documents that satisfy their information needs. The standard procedure is akin to looking for needles in a needlestack - the problem isn't so much that the desired information is not known, but rather that the desired information coexists with many other valid pieces of information. Just because a user is currently interested in NAFTA and not Furbies does not mean that all descriptions of Furbies are worthless. The problem is one of homing in on what is currently of interest to the user.

As noted above, the goal of data mining is to discover or derive new information from data, finding patterns across datasets, and/or separating signal from noise. The fact that an information retrieval system can return a document that contains the information a user requested does not imply that a new discovery has been made: the information had to have already been known to the author of the text; otherwise the author could not have written it down. I have observed that many people, when asked about text data mining, assume it should have something to do with ``making things easier to find on the web". For example, the description of the KDD-97 panel on Data Mining and the Web stated. Two challenges are predominant for data mining on the Web [3]. The first goal is to help users in finding useful information on the Web and in discovering knowledge about a domain that is represented by a collection of Web-documents. The second goal is to analyse the transactions run in a Web-based system, be it to optimize the system or to find information about the clients using the system

This search-centric view misses the point that we might actually want to treat the information in the web as a large knowledge base from which we can extract new, never-before encountered information.

On the other hand, the results of certain types of text processing can yield *tools* that indirectly *aid* in the information access process. Examples include text clustering to create thematic overviews of text collections automatically generating term associations to aid in query expansion and using co-citation analysis to find general topics within a collection or identify central web pages Aside from providing tools to aid in the standard information access process, I think text data mining can contribute along another dimension. In future I hope to see information access systems supplemented with tools for exploratory data analysis. Our efforts in this direction are embodied in the LINDI project, described in Section.

III. TDM AND COMPUTATIONAL LINGUISTICS

If we extrapolate from data mining (as practiced) on numerical data to data mining from text collections, we discover that there already exists a field engaged in text data mining: corpus-based computational linguistics[4] Empirical computational linguistics computes statistics over large text collections in order to discover useful patterns. These patterns are used to inform algorithms for various sub problems within natural language processing, such as part-of-speech tagging, word sense disambiguation, and bilingual dictionary creation

It is certainly of interest to a computational linguist that the words ``prices, prescription, and patent" are highly likely to co-occur with the medical sense of ``drug" while ``abuse, paraphernalia, and illicit" are likely to co-occur with the illegal drug sense of this word However, the kinds of patterns found and used in computational linguistics are not likely to be what the general business community hopes for when they use the term text data mining.

Within the computational linguistics framework, efforts in automatic augmentation of existing lexical structures seem to fit the data-mining-as-ore-extraction metaphor. Examples include automatic augmentation of WordNet relations by identifying lexico-syntactic [4] patterns that unambiguously indicate those relations, and automatic acquisition of subcategorization data from large text corpora. However, these serve the specific needs of computational linguistics and are not applicable to a broader audience.

IV. TDM AND CATEGORY METADATA

Some researchers have claimed that text categorization should be considered text data mining. Although analogies can be found in the data mining literature (e.g., referring to classification of astronomical phenomena as data mining, I believe when applied to text categorization this is a misnomer. Text categorization is a boiling down of the specific content of a document into one (or more) of a set of pre-defined labels. This does not lead to discovery of new information; presumably the person who wrote the document knew what it was about. Rather, it produces a compact summary of something that is already known.

However, there are two recent areas of inquiry that make use of text categorization and do seem to fit within the conceptual framework of discovery of trends and patterns within textual data[5] for more general purpose usage.

One body of work uses text category labels (associated with Reuters newswire) to find unexpected patterns" among text articles. The main approach is to compare distributions of category assignments within subsets of the document collection. For instance, distributions of commodities in country C1 are compared against those of country C2 to see if interesting or unexpected trends can be found. Extending this idea, one country's export trends might be compared against those of a set of countries that are seen as an economic unit.

Another effort is that of the DARPA Topic Detection and Tracking initiative.. While several of the tasks within this initiative are standard text analysis problems (such as categorization and segmentation), there is an interesting task called On-line New Event Detection, whose input is a stream of news stories in chronological order, and whose output is a yes/no decision for each story, made at the time the story arrives, indicating whether the story is the first reference[6] to a newly occurring event. In other words, the system must detect the first instance of what will become a series of reports on some important topic. Although this can be viewed as a standard classification task (where the class is a binary assignment to the new-event class) it is more in the spirit of data mining, in that the focus is on discovery of the beginning of a new theme or trend.

The reason I consider this examples - using multiple occurrences of text categories to detect trends or patterns - to be ``real" data mining is that they use text metadata to tell us something about the world, outside of the text collection itself. (However, since these applications use metadata associated with text documents, rather than the text directly, it is unclear if it should be considered text data mining or standard data mining.) The computational linguistics applications tell us about how to improve language analysis, but they do not discover more widely usable information.

The various contrasts made above are summarized in Table 1.

Table: 1 A classification of data mining and text data mining applications.

	Finding Patterns	**F inding Nuggets**	
		Novel	Non-Novel
Non-textual data	standard data mining	?	database queries
Textual data	computational linguistics	real TDM	information retrieval

Text Data Mining as Exploratory Data Analysis

Another way to view text data mining is as a process of exploratory data analysis that leads to the discovery of heretofore unknown information, or to answers to questions for which the answer is not currently known.

Of course, it can be argued that the standard practice of reading textbooks, journal articles and other documents helps researchers in the discovery of new information, since this is an integral part of the research

process. However, the idea here is to use text for discovery in a more direct manner. Two examples are described below.

V. USING TEXT TO FORM HYPOTHESES ABOUT DISEASE

For more than a decade, Don Swanson has eloquently argued why it is plausible to expect new information to be derivable from text collections: experts can only read a small subset of what is published in their fields and are often unaware of developments in related fields. Thus it should be possible to find useful linkages between information in related literatures, if the authors of those literatures rarely refer to one another's work. Swanson has shown how chains of causal implication within the medical literature[7] can lead to hypotheses for causes of rare diseases, some of which have received supporting experimental evidence.

For example, when investigating causes of migraine headaches, he extracted various pieces of evidence from titles of articles in the biomedical literature. Some of these clues can be paraphrased as follows:

- stress is associated with migraines
- stress can lead to loss of magnesium
- calcium channel blockers prevent some migraines
- magnesium is a natural calcium channel blocker
- spreading cortical depression (SCD) is implicated in some migraines
- high leveles of magnesium inhibit SCD
- migraine patients have high platelet aggregability
- magnesium can suppress platelet aggregability

These clues suggest that magnesium deficiency may play a role in some kinds of migraine headache; a hypothesis which did not exist in the literature at the time Swanson found these links. The hypothesis has to be tested via non-textual means, but the important point is that a new, potentially plausible medical hypothesis was derived from a combination of text fragments and the explorer's medical expertise. (According to swanson91, subsequent study found support for the magnesium-migraine hypothesis.

This approach has been only partially automated. There is, of course, a potential for combinatorial explosion of potentially valid links. beeferman98 has developed a flexible interface and analysis tool for exploring certain kinds of chains of links among lexical relations within WordNet[8] However, sophisticated new algorithms are needed for helping in the pruning process, since a good pruning algorithm will want to take into account various kinds of semantic constraints. This may be an interesting area of investigation for computational linguists.

VI. USING TEXT TO UNCOVER SOCIAL IMPACT

Switching to an entirely different domain, consider a recent effort to determine the effects of publicly financed research on industrial advances. After years of preliminary studies and building special purpose tools, the authors found that the technology industry relies more heavily than ever on government-sponsored research results. The authors explored relationships among patent text and the published research literature, using a procedure which was reported as follows in broad97:

The CHI Research team examined the science references on the front pages of American patents in two recent periods - 1987 and 1988, as well as 1993 and 1994 - looking at all the 397,660 patents issued. It found 242,000 identifiable science references and zeroed in on those published in the preceding 11 years, which turned out to be 80 percent of them. Searches of computer databases allowed the linking of 109,000 of these references to known journals and authors' addresses. After eliminating redundant citations to the same paper, as well as articles with no known American author, the study had a core collection of 45,000 papers. Armies of aides then fanned out to libraries to look up the papers and examine their closing lines, which often say who financed the research. That detective work revealed an extensive reliance on publicly financed science.

Further narrowing its focus, the study set aside patents given to schools and governments and zeroed in on those awarded to industry. For 2,841 patents issued in 1993 and 1994, it examined the peak year of literature references, 1988, and found 5,217 citations to science papers.

Of these, it found that 73.3 percent had been written at public institutions - universities, government labs and other public agencies, both in the United States and abroad.

Thus a heterogeneous mix of operations was required to conduct a complex analyses over large text collections. These operations included:

1. Retrieval of articles from a particular collection (patents) within a particular date range.
2. Identification of the citation pool (articles cited by the patents).
3. Bracketing of this pool by date, creating a new subset of articles.
4. Computation of the percentage of articles that remain after bracketing.
5. Joining these results with those of other collections to identify the publishers of articles in the pool.
6. Elimination of redundant articles.
7. Elimination of articles based on an attribute type (author nationality).
8. Location of full-text versions of the articles.
9. Extraction of a special attribute from the full text (the acknowledgement of funding).
10. Classification of this attribute (by institution type).
11. Narrowing the set of articles to consider by an attribute (institution type).
12. Computation of statistics over one of the attributes (peak year)
13. Computation of the percentage of articles for which one attribute has been assigned another attribute type (whose citation attribute has a particular institution attribute).

Because all the data was not available online, much of the work had to be done by hand, and special purpose tools were required to perform the operations.

VII.THE LINDI PROJECT

The objectives of the LINDI project[9] are to investigate how researchers can use large text collections in the discovery of new important information, and to build software systems to help support this process. The main tools for discovering new information are of two types: support for issuing sequences of queries and related operations across text collections, and tightly coupled statistical and visualization tools for the examination of associations among concepts that co-occur within the retrieved documents. Both sets of tools make use of attributes associated specifically with text collections and their metadata. Thus the broadening, narrowing, and linking of relations seen in the patent example should be tightly integrated with analysis and interpretation tools as needed in the biomedical example.

Following amant96, the interaction paradigm is that of a mixed-initiative balance of control between user and system. The interaction is a cycle in which the system suggests hypotheses and strategies for investigating these hypotheses, and the user either uses or ignores these suggestions and decides on the next move.

We are interested in an important problem in molecular biology that of automating the discovery of the function of newly sequenced genes. Human genome researchers perform experiments in which they analyze co-expression of tens of thousands of novel and known genes simultaneously[10] Given this huge collection of genetic information, the goal is to determine which of the novel genes are medically interesting, meaning that they are co-expressed with already understood genes which are known to be involved in disease. Our strategy is to explore the biomedical literature, trying to formulate plausible hypotheses about which genes are of interest.

Most information access systems require the user to execute and keep track of tactical moves, often distracting from the thought-intensive aspects of the problem. The LINDI interface provides a facility for users to build and so reuse sequences of query operations via a drag-and-drop interface. These allow the user to repeat the same sequence of actions for different queries. In the gene example, this allows the user to specify a sequence of operations to apply to one co-expressed gene, and then iterate this sequence over a list of other co-expressed genes that can be dragged onto the template. (The Visage interface. implements this kind of functionality within its information-centric framework.) These include the following operations.

Iteration of an operation over the items within a set. (This allows each item retrieved in a previous query to be use as a search terms for a new query.)

- Transformation, i.e., applying an operation to an item and returning a transformed item (such as extracting a feature).
- Ranking, i.e., applying an operation to a set of items and returning a (possibly) reordered set of items with the same cardinality.
- Selection, i.e., applying an operation to a set of items and returning a (possibly) reordered set of items with the same or smaller cardinality.
- Reduction, i.e., applying an operation to one or more sets of items to yield a singleton result (e.g., to compute percentages and averages).

This system will allow maintenance of several different types of history including history of commands issued, history of strategies employed, and history of hypotheses tested. For the history view, we plan to use a spreadsheet" layout as well as a variation on a ``slide sorter" view which Visage uses for presentation creation but not for history retention.

Since gene function discovery is a new area, there is not yet a known set of exploration strategies. So initially the system must help an expert user generate and record good exploration strategies. The user interface provides a mechanism for recording and modifying sequences of actions. These include facilities that refer to metadata structure, allowing, for example, query terms to be expanded by terms one level above or below them in a subject hierarchy. Once a successful set of strategies has been devised, they can be re-used by other researchers and (with luck) by an automated version of the system. The intent is to build up enough strategies that the system will begin to be used as an assistant or advisor., ranking hypotheses according to projected importance and plausibility.

Thus the emphasis of this system is to help automate the tedious parts of the text manipulation process and to integrate underlying computationally-driven text analysis with human-guided decision making within exploratory data analysis over text.

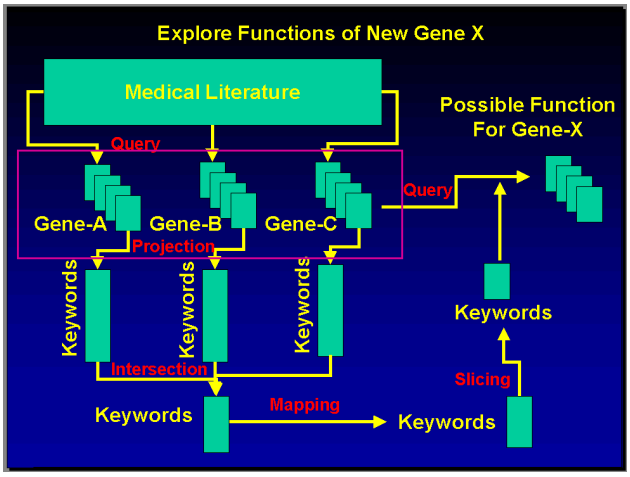

Figure: 1 A hypothetical sequence of operations for the exploration of gene function within a biomedical text collection, where the functions of genes A, B, and C are known, and commonalities are sought to hypothesize the function of the unknown gene. The mapping operation imposes a rank ordering on the selected keywords. The final operation is a selection of only those documents that contain at least one of the top-ranked keywords and that contain mentions of all three known genes.

VIII. CONCLUSION

For almost a decade the computational linguistics community has viewed large text collections as a resource to be tapped in order to produce better text analysis algorithms. In this paper, I have attempted to suggest a new emphasis: the use of large online text collections to discover new facts and trends about the world itself. I suggest that to make progress we do not need fully artificial intelligent text analysis; rather, a mixture of computationally-driven and user-guided analysis may open the door to exciting new results.

REFERENCES

1. Armstrong, Susan, (Ed.) (1994) "Using Large Corpora". MIT Press.
2. Berry, Michael & Linoff, Gordon (1997) "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons.
3. Cherkassky V. & Mulier, F. (1998) "Learning From Data", John Wiley & Sons.
4. Fayyad, Usama; Haussler, David & Stolorz, Paul (1996) "Mining Scientific Data", Communications of the ACM, vol. 39, no. 11, pp. 51-57, November 1996
5. Fieldman, Ronen & Sanger, James (2007) "The Text Mining Handbook, Advanced Approaches in Analysing Unstructured Data" Cambridge University Press.
6. Hand, David; Mannila, Heikki & Smyth, Padhraic (2001) "Principles of Data Mining" The MIT Press
7. Hastie T., Tibshirani R. & Friedman J. (2001) "The Elements of Statistical Learning: Data Mining, Inference and Prediction", Springer-Verlag
8. Hearst, Marti A. (1997) "Text Data Mining, Issues, Techniques, and the Relationship to Information Access" UW/MS Workshop on Data Mining, July 1997, University of Berkley, July, 1997.
9. Ronen Feldman, Will Klosgen, and Amir Zilberstein. 1997. Visualization techniques to explore data mining results for document collections. In *Proceedings of the Third Annual Conference on Knowledge Discovery and Data Mining (KDD)*, Newport Beach.
10. David C. Hoaglin, Frederick Mosteller, and John W. Tukey. 1983. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, Inc.

