Journal of Computer Science and Applications.
ISSN 2231-1270 Volume 5, Number 2 (2013), pp. 51-55

http://www.irphouse.com

Text Mining - Scope and Applications

Miss Latika Kaushik

Software Developer at ACSG, Delhi

Abstract

Text mining which is also known as data mining from textual unstructured databases refers to the process of extracting interesting and non-trivial patterns or knowledge from text i.e. may it be in any format. Unstructured text is very common, and in fact may represent the majority of information available to a particular research or data mining project. Has text mining evolved so rapidly to become a mature field? This article attempts to shed some lights to the scope and applications of Text Mining. In conclusion, we highlight the scope , upcoming challenges of text mining and the opportunities it offers.

What is text mining ?

Text mining is an art of “text analytics” which is one way to make qualitative or “unstructured” data usable by a computer. Qualitative data is descriptive data that cannot be measured in numbers and often includes qualities of appearance like color, texture, and textual description. Quantitative data is numerical, structured data that can be measured. However, there is often confusion between qualitative and quantitative categories. For example, a photograph might traditionally be considered “qualitative data” but when you break it down to the level of pixels, which can be measured.

Text mining can help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Facebook, Twitter and professional media like LinkedIn. Mining unstructured data with natural language processing (NLP), statistical modeling and machine learning techniques can be challenging, however, because natural language text is often inconsistent. It contains ambiguities caused by inconsistent syntax and semantics, including slang, language specific to vertical industries and age groups, double entendres and sarcasm.

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-

quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

Why do Text Mining?

Broadly speaking there are (so far) four main reasons for text mining:

1. to enrich the content in some way
2. to enable systematic review of literature
3. for discovery
4. for computational linguistics research.

Text Mining and Data Mining

Just as data mining can be loosely described as looking for patterns in data, text mining is about looking for patterns in text. However, the superficial similarity between the two conceals real differences. Data mining can be more fully characterized as the extraction of implicit, previously unknown, and potentially useful information from data .The information is implicit in the input data: it is hidden, unknown, and could hardly be extracted without recourse to automatic techniques of data mining. With text mining, however, the information to be extracted is clearly and explicitly stated in the text. It's not hidden at all—most authors go to great pains to make sure that they express themselves clearly and unambiguously—and, from a human point of view, the only sense in which it is "previously unknown" is that human resource restrictions make it infeasible for people to read the text themselves.

The problem, of course, is that the information is not couched in a manner that is amenable to automatic processing. Text mining strives to bring it out of the text in a form that is suitable for consumption by computers directly, with no need for a human intermediary.

Text Mining Tools

Computer program makes any task easier .Text mining computer programs are available from many commercial and open source companies and sources.

Commercial

- AeroText – a suite of text mining applications for content analysis. Content used can be in multiple languages.
- Angoss – Angoss Text Analytics provides entity and theme extraction, topic categorization, sentiment analysis and document summarization capabilities via the embedded Lexalytics Salience Engine. The software provides the unique capability of merging the output of unstructured, text-based analysis with structured data to provide additional predictive variables for improved predictive models and association analysis.

- Attensity – hosted, integrated and stand-alone text mining (analytics) software that uses natural language processing technology to address collective intelligence in social media and forums; the voice of the customer in surveys and emails; customer relationship management; e-services; research and e-discovery; risk and compliance; and intelligence analysis.
- Autonomy – text mining, clustering and categorization software

Lots many open source tools are also available.

Practical Applications of Text Mining

Some examples of practical applications of text mining techniques include:

- Spam filtering
- Creating suggestion and recommendations (like amazon)
- Monitoring public opinions (for example in blogs or review sites)
- Customer service, email support
- Automatic labeling of documents in business libraries
- Measuring customer preferences by analyzing qualitative interviews
- Fraud detection by investigating notification of claims
- Fighting cyberbullying or cybercrime in IM and IRC chat

and so on…

Let me discuss few of them-

Security applications

Many text mining software packages are marketed for security applications, especially monitoring and analysis of online plain text sources such as Internet news, blogs, etc. for national security purposes. It is also involved in the study of text encryption/decryption.

Software applications

Text mining methods and software is also being researched and developed by major firms, including IBM and Microsoft, to further automate the mining and analysis processes, and by different firms working in the area of search and indexing in general as a way to improve their results. Within public sector much effort has been concentrated on creating software for tracking and monitoring terrorist activities.

Online media applications

Text mining is being used by large media companies, such as the Tribune Company, to clarify information and to provide readers with greater search experiences, which in

turn increases site "stickiness" and revenue. Additionally, on the back end, editors are benefiting by being able to share, associate and package news across properties, significantly increasing opportunities to monetize content.

Marketing applications

Text mining is starting to be used in marketing as well, more specifically in analytical customer relationship management. A survey and marketing is done by using many of Text Mining Tools.

Sentiment analysis

Sentiment analysis may involve analysis of movie reviews for estimating how favorable a review is for a movie.Such an analysis may need a labeled data set or labeling of the affectivity of words. Text has been used to detect emotions in the related area of affective computing.Text based approaches to affective computing have been used on multiple corpora such as students evaluations, children stories and news stories.

Academic applications

The issue of text mining is of importance to publishers who hold large databases of information needing indexing for retrieval. This is especially true in scientific disciplines, in which highly specific information is often contained within written text.

Potential Weakness:

Finally, as a word of caution, text mining doesn't generate new facts and is not an end, in and of itself. The process is most useful when the data it generates can be further analyzed by a domain expert, who can bring additional knowledge for a more complete picture. Still, text mining creates new relationships and hypotheses for experts to explore further.

Challenges of Text Mining

There exists lots of challenges for data mining .Let me discuss few of them-

1) Very high number of possible "dimensions"
– All possible word and phrase types in the language!
2) Unlike data mining:
– records (= docs) are not structurally identical
– records are not statistically independent
3) Complex and subtle relationships between concepts in text
– "POL merges with Time-Changer"

– "Time-Changer is bought by POL"
4) Ambiguity and context sensitivity
– automobile = car = vehicle = Toyota
5) Data collection is "free text"
– Data is not well-organized
6) Semi-structured or unstructured
– Natural language text contains ambiguities on many levels
7) Lexical, syntactic, semantic, and pragmatic
– Learning techniques for processing text typically need annotated training examples

CONCLUSION

We have provided a very brief introduction to text mining , its scope and applications within the pages of this article. There are numerous challenges to the statistical community that reside within this discipline area. As in any data mining or exploratory data analysis effort, visualization of textual data is an essential part of the problem. The statistical community has a great deal to contribute to many of these problems. In this article we have presented the scope and challenges in the area of text mining. .

References

[1] Wikipedia , http://en.wikipedia.org/wiki/Text_mining
[2] http://eprints.ncrm.ac.uk/227/1/What_is_Text_Mining.pdf
[3] Text Mining: Applications and Theory , Wiley InterScience
[4] http://www.statsoft.com/textbook/text-mining/
[5] http://searchbusinessanalytics.techtarget.com/definition/text-mining
[6] http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf
[7] http://www.publishingresearch.net/documents/PRCTextMiningandScholarlyPublishinFeb2013.pdf
[8] http://text-analysis.sourceforge.net/practical-applications
[9] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (kdt)," in Proceedings of the Conference on Knowledge Discovery and Data Mining, 1995,
[10] Google Search Engine