Asian Journal of Information Technology 13 (2): 93-98, 2014
ISSN: 1682-3915

A New Text Mining Approach in Search Technology

[1]B. SunilSrinivas, [2]P.N. Santosh Kumar, [3]A. Govardhan and [2]C. Sunil Kumar
[1]Department of CSE, VVIT, Hyderabad, A.P., India
[2]Department of ECM, SNIST, Hyderabad, A.P., India
[3]School of IT, JNTUH, Hyderabad, A.P., India

Abstract: Text-Mining (TM) refers generally to the practice of extracting attractive and non-trivial information and facts from unstructured text. TM includes several Computer Science (CS) regulations with a strong direction towards Artificial Intelligence (AI) in general including but not limited to Pattern Recognition (PR), Neural Networks (NN), Natural Language Processing (NLP), Information Retrieval (IR) and Machine Learning (ML). A significant variation with search is that search requires a user to identify what he or she is looking for while TM attempts to realize information in a model that is not known earlier. TM is mainly motivating in domains where users have to invent new information. This is the case for, e.g., in criminal enquiries and legal findings. Such examinations require 100% evoke, i.e., users can not meet the expense of missing relevant data. In distinction, a user searching the internet for background information using a benchmark Search Engine (SE) simply requires any data as long as it is reliable. Increasing evoke almost positively will decrease accuracy involving that users have to browse huge collections of documents that that may or may not be relevant. Standard procedures use language expertise to increase accuracy but when text collections are not in one language are not domain specific and or contain variable size and type documents either these schemes fail or are so complicated that the user does not understand what is happening and loses control. A different technique is to combine standard significance ranking with Adaptive Filtering (AF) and Interactive Visualization (IV) that is based on characteristics that have been mined earlier.

Key words: TM, AI, PR, DM, NLP, PRA, SE

INTRODUCTION

Within the area of expertise of TM, now and then called text analytics, several attractive technologies such as computers, IT, PR, statistics, advanced mathematical techniques, AI, visualization and IR. The information bang of modern times will persist at the same rate (Allan, 2002; Andrews, 2008). TM techniques play a crucial role in the upcoming years in this lifelong process. Due to continuing globalization there is also much interest in Multi-Language Text Mining (MLTM) TM: the attaining of insights in ML collections. MLTM is much more difficult that it appears as in addition to differences in words and character sets, TM makes concentrated use of data as well as the linguistic properties of a language. There are many essential hypotheses about capitalization and tokenization that would not work for other languages. When TM Methods are used on non-English data collections supplementary challenges have to be concentrated (Berry, 2004; Berry and Castellanos, 2006).

TM is about investigating unstructured information and extracting relevant patterns or models and uniqueness. Using these models and characteristics better search results and deeper data analysis is possible; giving quick IR otherwise it would remain hidden. The field of Data Mining (DM) is better known than that of TM. A good, e.g., of DM is the analyzing of operational details contained in relational databases, such as debit card transactions or credit card payments. To such operational diverse supplementary information can be provided: date, location, age of card holder, salary, etc. With support of this information patterns of behavior can be determined.

However, 95% of all information is unstructured information and both the proportion and the total amount of unstructured information raise daily. Only a small amount of information is stored in a structured format in a relational database. The greater part of information that users work on every day is in the form of Text Documents (TD), e-mails/multimedia files (speech, video and photos). Searching analysis using Database (DB) or DM Methods of this information is not possible as these procedures

Corresponding Author: B. SunilSrinivas, Department of CSE, VVIT, Hyderabad, A.P., India

work only on Structured Information (SI). It is easier to manage, share, search, organize and to generate reports so on for computers as well as users therefore the wish is to give structure to unstructured information. This allows computers and public to better manage the data and allow known procedures and methods to be used.

TM, using manual procedures was use first during the 1980s. It hastily became obvious that these manual procedures were labor demanding and therefore costly (Bilisoly, 2008). It also cost too much time to manually process the already-growing amount of information. Over time there was growing success in creating applications to mechanically process the information and in the last 10 years there has been much development.

Currently, the study of TM worries the growth of various mathematical, statistical and methods which allow mechanical analysis of unstructured information as well as the extraction of high quality and appropriate data and to make the text as a whole better searchable. High excellence refers here in particular to the combination of the relevance and the obtaining of new and interesting approaches. A TD contains characters that together form words which can be combined to form phrases. These are all syntactic properties that collectively symbolize defined categories, concepts, meanings. TM must distinguish, extract and use all this information. Using TM, instead of searching for words, researchers can search for linguistic (scientific study of language) word patterns and this is therefore searching at a higher level.

SEARCHING UNSTRUCTURED INFORMATION

What turns out exactly when somebody uses a computer program to search unstructured text? Computers are digital apparatus with limited capabilities. Computers manage best with numbers, in particular whole numbers also known as integers, if it has to be really fast. Public are analogue and individual language is also analogue, full of irregularity, obstruction, errors and exceptions. If public search for something then they often think in concepts and meanings, all areas in which a computer can not directly transact with.

For machines to be able to make a computationally proficient search in a huge amount of text, the difficulty needs first to be transformed to a numerical problem that a computer can associate with. This directs to very large storages containing many numbers in which numbers characterizing search terms are compared with numbers characterizing information and documents. This is the basic standard that the field concerns itself with: translation of information that users can work with into information that a computer can work with and then convert the result back into a form that public can understand.

This expertise exists since the 1960s. One of the first scientists working in this field was Gerard Salton, who together with others made one of the first texts SE. Each incident of a word in the text was entered in a keyword index. Searching was then done in the index, analogous to the index at the back of a book but with many more words and much earlier. With procedures such as B-trees and hashing, it was probable to speedily and proficiently make a list from all documents containing a word or a Boolean combination of words (Blair and Maron, 1985).

Documents and search terms were transformed to vectors and contrasted using the cosine distance among them: how lesser the cosine distance, how additional the search term and the document communicated. This was an efficient technique to decide the relevance of a document from the search expression. This was called the vector space model and is still used today by some applications. Later on, a variety of other techniques used for searching and relevance. There are many search procedures with good-sounding names such as: directed and non-directed proximity, fuzzy, wildcards, semantical, taxonomies, conceptual, etc. Examples of normally known relevance defining procedures are: Term-based Frequency Ranking (TBFR), the Page-Rank Algorithm (PRA) and Probabilistic Ranking (PR).

Because these days there is so much information digitally available and because it is now often crucial to directly react on present activities, new procedures are necessary to keep up with the continuously rising amount of unstructured information. In addition, people will have different causes for searching large amount of data and different goals to find and those distinctions require an alternative advances.

SEARCHING AND FINDING

Fraud researchers or public prosecutors don't only want the best documents; they want all probable appropriate documents. In addition, in an internet SE everyone does their best to get to the top of the results list; search engine optimization has in itself become a skill. This is done by using synonyms and code names and relatively often these are common words that are used so often that a search cannot be done without returning millions of knocks (hits). TM can offer a solution to find the appropriate information. Fraud researchers also have another frequent difficulty; at the beginning of the

analysis they do no know accurately what they must search for. They do not know the code names or synonyms or they do not exactly know which companies, account numbers, persons or amounts must be searched for. Using TM it is likely to recognize all these types of properties from their linguistic role and then to categorize them in a structured manner to present them to the user. It then becomes very simple to investigate the found companies further.

Sometimes the difficulties facing by a researcher go a little deeper; they are searching without really knowing what they are searching for. TM can be used to discover the words and subjects important for the examination; the computer searches for specified patterns in the text: "who paid to whom" "who talked to whom", etc. These types of patterns can be predictable using language technology and TM and extracted from the text and presented to the researcher who can then quickly decide the legal transactions from the suspect ones.

An example; if the HDFC bank transfers money to the AXIS then that is a normal transaction. But if "Norman" transfers money to Kumaran Enterprises Inc. then that may be doubtful. TM can recognize these natures of patterns and further searches can be made on the words in those patterns using normal search procedures to further recognize and examine details. The obtaining of new insights is also called fortune. TM can be custom-made very efficiently to obtain new but often vital insights necessary to progress in an examination.

Therefore, the TM helps in the search for data by using models for which the values of the elements are not exactly known earlier. This is analogous with mathematical functions in which the variables and the statistical distribution of the variables are not always known. Here, the heart of the difficulty can be seen as a conversion problem from human language to mathematics. The better the mathematical conversion, the better excellence of the TM will be.

INFORMATION VISUALIZATION

TM is often stated as Information Visualisation (IV). This is because visualisation is one of the practical possibilities after unstructured information has been structured. To be able to make these sorts of visualizations the features must be structured or planned and that is accurately the area in which TM expertise can help: by structuring unstructured information it is possible to visualise the data and more rapidly obtain new insights (Card *et al*., 1999).

> ZyLAB donates a PACS to the Government of US Washington, The Pacific Northwest, July 16th, 2001-ZyLAB, the developer of document imaging and full-text retrieval software has donated a PACS System to the government of US (Baron, 2005). "We have been working closely with the World Health Organization, US International for the last 4 years now," said, CEO of ZyLAB Technologies. "The US government faced difficult task with the ZyLAB System will be of tremendous assistance to them. Unfortunately, the US has scarce resources to procure advanced imaging and archiving systems to help them in this task, so we decided to donate them a full operational PAC System." A demonstration of the ZyLAB Software was done for the US by David of the Criminal Justice Resource Center, an American-Canadian volunteer group: "The US was greatly impressed. They want and need this system as they currently have evidence sitting in folders that is difficult to search. This is one of the major delays in getting the 1,20,000 accused persons in custody to trial." My hope and belief is that ZylMAGE will enable Mr. Gahima's office to process, preserve and catalogue the US evidence collection so that the significance and details of the genocide in US can be preserved"

An example is the following text; in that text, the following entities and attributes can be found in the Table 1. Let's assume that there are various documents containing this type of automatically found structured properties then the documents could not only be presented in tabular form but also for, e.g., in a tree structure in which the document could be structured on occurrences per land and then on occurrences per administration. The principle can also be used to dynamically imagine a tree structure which would then appear as shown in Fig. 1. These types of visualization methods are ideal for allowing an easy insight into large e-mail collections. Alongside the structure that TM methods can deliver use can also be made of the available

Table 1: Entities and attributes

Places	Washington
Countries	The Pacific Northwest, US
Persons	Gerald Gahima, David
Function titles	CEO
Data	July 16th, 2001
Organisations	Government of US, WHO, Criminal Justice Resource Center, American-Canadian volunteer group
Companies	ZyLAB, ZyLAB Technologies BV
Products	PACS

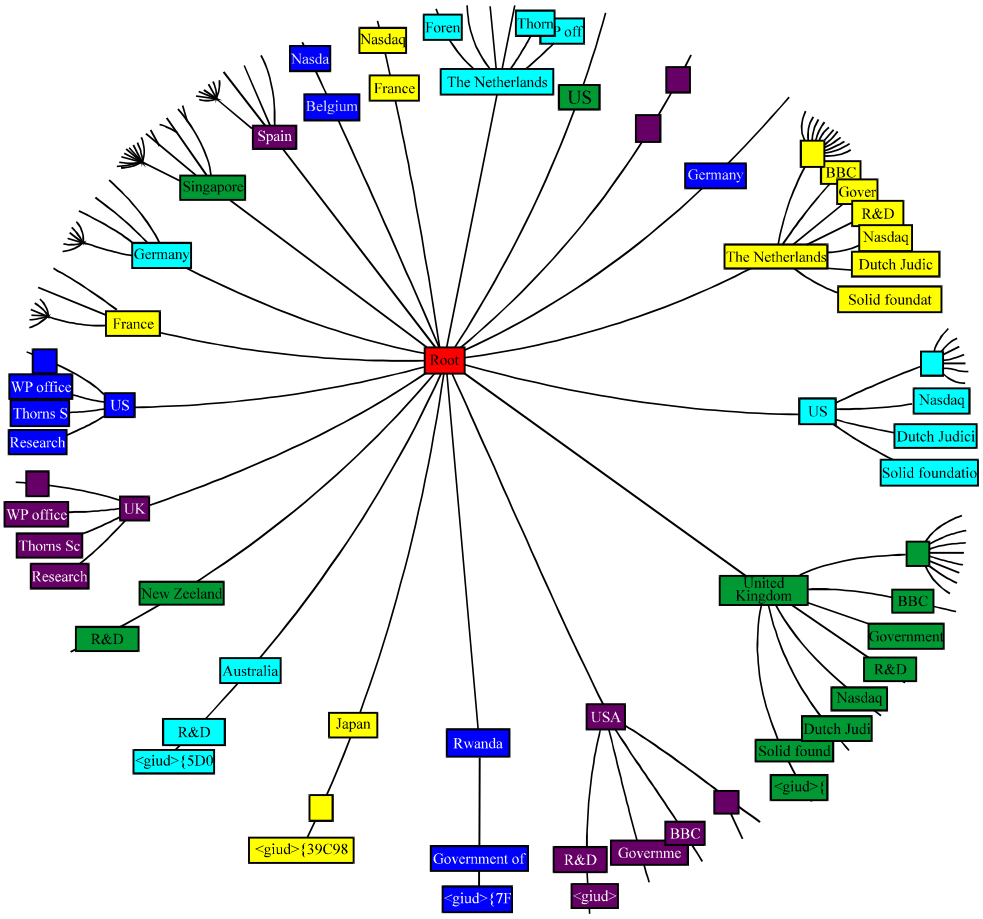

Fig. 1: Hyperbolic tree visualisation of a tree structure

attributes such as “Sender”, “Recipient”, “Subject”, “Date”, etc. Below a number of possibilities for e-mail visualization shown in the Fig. 2.

With the help from these types of visualisation methods it is possible to gain a quicker and better insight into difficult data collections, particularly if it involves large groups of unstructured information that can be mechanically structured using DM.

Advantages of structured and analysed data: In addition to the visualization, various other search extensions are possible when the data has been structured and has Meta details. Here is a concise list:

- Details are easier to organize in folders
- It is easier to filter data on specified Meta details when searching
- Details can be contrasted and linked using the Meta details
- It is probable to sort, group and prioritize the documents using any of the attributes
- Details can be clustered using the Meta details
- With the assist of Meta details duplicates and almost-duplicates can be sensed. These can then be deleted
- Taxonomies can be derived from the Meta details
- Rule-based analyses can be made on the Meta details
- It is feasible to search the Meta details from already found documents
- Statistical reports can be made on the basis of the Meta details
- It is possible to search for relationships between Meta details for, e.g., “who paid to whom and how much” in which the “who” and the “how much” are not previously known

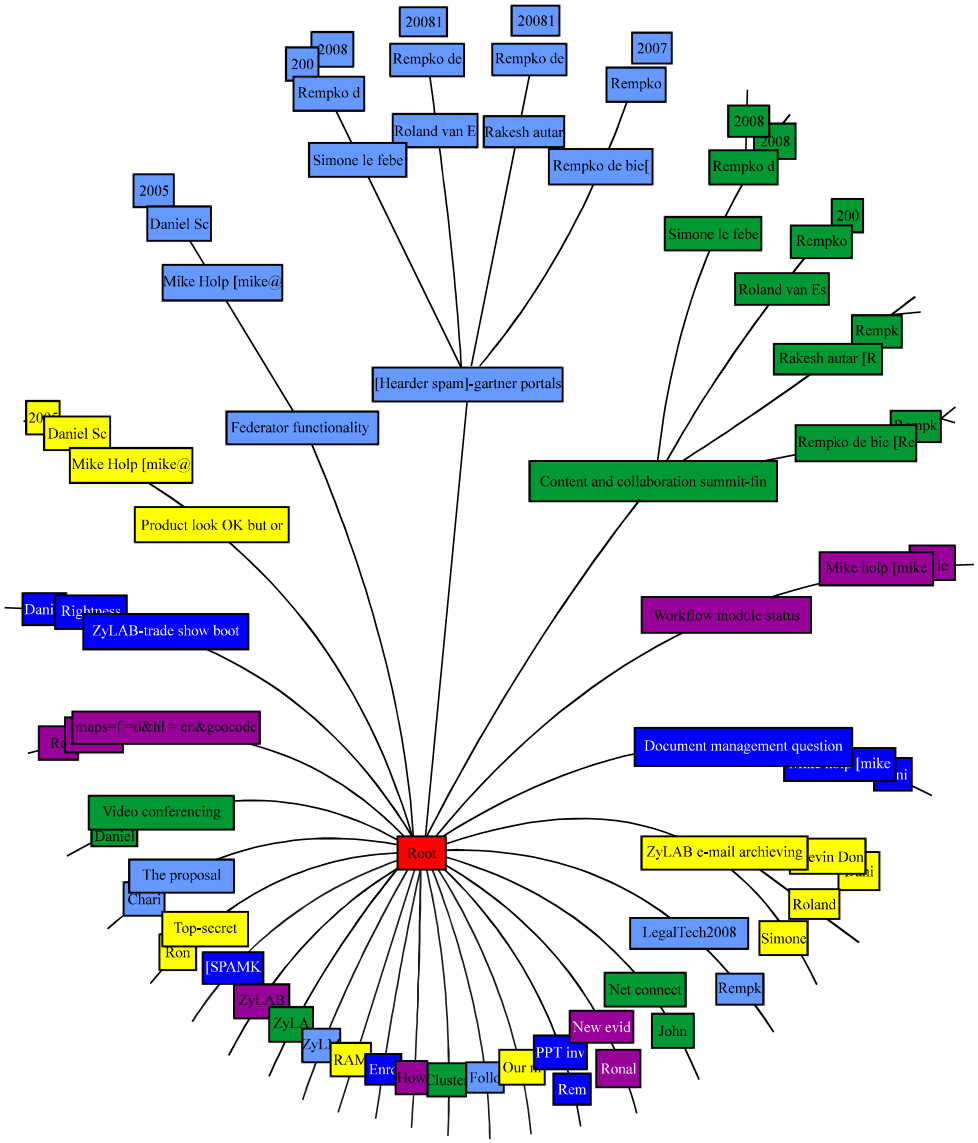

Fig. 2: E-mail visualisation using a hyperbolic tree

CONCLUSION

Although, changes in the authorized world are always evolutions and never revolutions, there is certainly a potential role for TM in e-Discovery and e-Disclosure. Data collections are just getting to large to be reviewed serially. Collections need to be pre-analysed and pre-organized. Re-examinations can be implemented more efficiently and goals can be made easier. The challenge will be to encourage courts of the rightness of these new tools.

Therefore, a hybrid advance is recommended where computers make the initial selection and classification of documents and research directions and human reviewers and researchers implement quality control and valuate the survey suggestions. By doing so, computers can focus on recall and human being can focus on accuracy. There are many other applications where this approach has led to both more competence but also to acceptance of the technology by the public.

REFERENCES

Allan, J., 2002. Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Publishers, Kluwer Academic Publishers.

Andrews, W., 2008. Magic quadrant for information access technology. Gartner Research Report, ID Number: G00161178. Gartner, Inc., September 30, 2008.

Baron, J.R., 2005. Toward a federal benchmarking standard for evaluating information retrieval products used in e-discovery. Sedona Conf. J., 6: 237-246.

Berry, M.W. and M. Castellanos, 2006. Survey of Text Mining II: Clustering, Classification and Retrieval. Springer-Verlag, Berlin, Germany.

Berry, M.W., 2004. Survey of Text Mining: Clustering, Classification and Retrieval. Springer-Verlag, Berlin, Germany.

Bilisoly, R., 2008. Practical Text Mining with Perl (Wiley Series on Methods and Applications in Data Mining). John Wiley and Sons, London, UK.

Blair, D.C. and M.E. Maron, 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. Commun. ACM, 28: 289-299.

Card, S.K. J.D. Mackinlay and B. Shneiderman, 1999. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann Publishers, San Francisco, CA., USA.