GLOBAL RESEARCH ANALYSIS

GLOBAL RESEARCH ANALYSIS

ISSN No 2277 - 8160

listed & Indexed with
International ISSN Directory, Paris



Multi-Subject Journal

Volume : 1 | Issue : 3 | August 2012 | ₹ 250/-

ISSN No 2277 - 8160

GLOBAL RESEARCH ANALYSIS

Editor In-Chief :

Dr. Khansa Memon
SARA Publishing Academy

Dr. Ashok S. Pawar Associate Professor, Dept. of Economic Dr. Babaasaheb Ambedkar Marathwada.University, Aurngabad.(MS-India)-431004	**Dr.(Prof) Vijay Kumar Soni** Principal, Jai Meenesh College, Phagi, Jaipur, Rajasthan	**Dr. A.R. Saravankumar** Assistant Professor in Education DDE, Alagappa University, Tamilnadu
Dr. R. Ramachandran Commerce DDE Annamalai University Tamilna	**Dr. R Ganpathy** Assistant Professor in Commerce Directorate of Distance Education Alagappa University Karaikudi.	**Dr. Amit Bandyopadhyay** Assistant Professor Department of Physiology University of Calcutta
Dr. V. Kumaravel Professor and Head Vivekanandha Business School for Women Tiruchengode, Namakkal Dist	**Dr. K. Prabhakar** Professor, Department of Management Studies, Velammal Engineering College, Chennai - 600 066.	**Dr. Sunita J. Rathod** Maharashtra Education Service Group-B DIET Dist. Jalna

SUBSCRIPTION DETAILS

Period	Rate	Discount	Amount Payable
One Year (12 issues)	3000/-	Nil	3000/-
Two Years (24 issues)	6000/-	200/-	5800/-
Three Years (36 issues)	9000/-	300/-	8700/-
Five Years (60 issues)	15000/-	600/-	14400/-

You can download the Subscription form from website www.theglobaljournals.com. You will require to print the form. Please fill the form completely and send it to the Editor, GLOBAL RESEARCH ANALYSIS along with the payment in the form of Demand Draft/Cheque at Par drawn in favour of GLOBAL RESEARCH ANALYSIS payable at Ahmedabad.

- Thoughts, language, vision and examples in published research paper are entirely of author of research paper. It is not necessary that both the editor, and editorial board agrees with the author. The responsibility of the matter of research paper/article is entirely of author.
- Editing of the Global Research Analysis is processed without any remittance. The selection and publication is done after recommendations of atleast two subject expert referees.
- In any condition if any National/International University denies accepting the research paper published in GRA, then it is not the responsibility of Editor, Publisher and Management
- Only the first author is entitle to receive the copies of all co-authors
- Before re-use of published paper in any manner, it is compulsory to take written permission from the Editor-GRA unless it will be assumed as disobedience of copyright rules
- All the legal undertaking related to Global Research Analysis is subject to Ahmedabad Jurisdiction.
- The research journal will be send by normal post. If the journal is not received by the author of research paper then it will not be the responsibility of the Editor and publisher. The amount for registered post should be borne by author of the research paper in case of second copy of the journal.

Editor,
GLOBAL RESEARCH ANALYSIS
8-A, BANAS, Opp. SLU Girls College, Nr.Congress Bhavan, Paldi, Ahmedabad - 380006. Gujarat (INDIA)
Cell: +91 9824702127, 8866003636

INDEX

Sr. No.	Title	Author	Subject	Page No.
1	Antipyretic Activity Of Tagetes Patula	K.Shakila, D.Sukumar, G.Madhubala	Chemistry	1-2
2	Benefits Derived From Warehouse Management System By Small Manufacturing Firms	Dr. Vipul Chalotra, Prof Neetu Andotra	Commerce	3-4
3	Women Welfare Through Prime Minister Rozghar Yojana At Coimbatore District	Swapna Kurian	Commerce	5-7
4	Sources Of College Funds In Assam-An Overview	Dr. Niranjan Kakati	Commerce	8-9
5	Mgnarega- For Socio-Economic Transformation Of Rural Society In J&K	TARSEM LAL	Commerce	10-11
6	Performance Appraisal And Career Growth	Dr. J.K. Sehgal	Commerce	12-13
7	Corporate Governance Practices - Emerging Trends	Dr. V.Rajendran	Commerce	14-16
8	A Text Mining Framework: Promises And Challenges	Prof.Sangeeta Mahesh Borde, Prof.Bareen Kayyum Shaikh	Computer Science	17-19
9	Effects Of Infant Mortalty Rates On Fertility	Mrs. Nazia Ansari, Mrs. Aqsa Khatri	Economics	20-21
10	Social Entrepreneurship Initiatives In India	Dr. Shradha H. Budhedeo, Parinda Doshi	Economics	22-25
11	Economics Of Corruption	Prof. Dr. K.J. Thankachan	Economics	26-27
12	A Study On Attitude Of High School Students Towards Physics Lessons And Physical Experiments	Dr. Kotreshwaraswamy A. Surapuramath	Education	28-30
13	Utilization Of Science Laboratories At High Schools	Dr. Praveena K. B.	Education	31-33
14	Mutidimensional Techniques For Mathematical Operations	M.Kavitha , Dr.AR. Saravanakumar	Education	34-36
15	Scientific Approach of Communication Theory	Dr. Mujibul Hasan Siddiqui	Education	37-38
16	Design And Implementation Of Ethernet Controller On Fpga	Jesika Benani	Engineering	39-41
17	Development of Control Algorithm to Reduce Harmonic in Boost Converter	Mitulkumar R. Dave, Prof. K.C. Dave	Engineering	42-43
18	Development Of Control Algorithm To Reduce Harmonics In Buck Converter	Vipul C. Rajyaguru, Keerti S.Vashishtha, K. C. Dave	Engineering	44-46
19	Gandhian Principles And Women Empowerment Through Self-Help Groups	Dr. Chetankumar N. Trivedi	English	47-48
20	Kathal- its medicinal uses	Dr. Sneh Harshendra Sharma	Environment	49-50
21	Extemporize Employee Engagement: Challenges.	Dr.Pooja Arora	Management	51-53
22	Conceptual Study To Understand The E-Governance Prestige Values Associated With Nagar Nigam	Dr.Murlidhar Panga , Mr.Arpit Neema	Management	54-56
23	Impact Of Globalization On Environment	Ankita Soni	Management	57-58
24	A Study On Strategic Cost Management Of Tanishq	Mr. Suresh A.S, Mr. Bharath R	Management	59-61

Sr. No.	Title	Author	Subject	Page No.
25	Impact Of Soap Operas Among The Housewives And Working Woman Of Silchar Town	Dr Partha Sarkar	Mass Communication	62-64
26	A Tool For Sex Identification	Dr. Renuka B. Adgaonkar, Dr. M. P. Ambali	Medical Science	65-67
27	Ideal characters in Sanskrit literature	Dr. Labhuben Maldebhai Karavdara	Sanskrit	68
28	Progresive life in vaidik life	Dr. Labhuben Maldebhai Karavdara	Sanskrit	69
29	Familisation or family in Vaidik Literature	Dr. Labhuben Maldebhai Karavdara	Sanskrit	70
30	Morden Moral Of Modern Environment In Nature's Description Of Vishnupuran	Dr. Labhuben Maldebhai Karavdara	Sanskrit	71-72
31	India's Small Holder Dairying In Liberalized Era	Dr Ashish.K.Makwana	Science	73-74
32	Common Man And The Uncommon Law : An Overview	Dr. Shereen Sadiq	Social Science	75-77
33	Perspectives Of Counselling	Dr, R.DhakshinaMurthy	Social Sciences	78-80
34	Sri Sri And His Literary Works: Some Social Reflections	Dr. M. Sreedevi Xavier	Social Sciences	81-83

Volume : 1 | Issue : 3 | Aug 2012 • ISSN No 2277 - 8160

Research Paper

Computer Science

A Text Mining Framework: Promises And Challenges

Prof.Sangeeta Mahesh Borde Department of Computer Science, MIT College of Arts, Commerce & Science,Alandi

Prof.Bareen Kayyum Shaikh Department of Computer Science, MIT College of Arts, Commerce & Science,Alandi

ABSTRACT

Text mining is a burgeoning new field that attempts to glean meaningful information from natural language text. It may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, text is the most common vehicle for the formal exchange of information. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling—even if success is only partial. Text mining, also known as knowledge discovery from text, and document information mining, refers to the process of extracting interesting patterns from very large text corpus for the purposes of discovering knowledge. Text mining is an interdisciplinary field involving information retrieval, text understanding, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining. Regarded by many as the next wave of knowledge discovery, text mining has a very high commercial value. This talk presents a general framework for text mining, consisting of two stages: text refining that transforms unstructured text documents into an intermediate form; and knowledge distillation that deduces patterns or knowledge from the intermediate form. In conclusion, we highlight the upcoming challenges of text mining and the opportunities it offers

KEYWORDS: Text Mining, knowledge, categorization, visualization

1. INTRODUCTION

Text mining aims to extract useful knowledge from textual data or documents (Hearst, 1999; Chen, 2001). Although text mining is often considered a subfield of data mining, some text mining techniques have originated from other disciplines, such as information retrieval, information visualization, computational linguistics, and information science. Examples of text mining applications include document classification, document clustering, entity extraction, information extraction, and summarization. As the most natural form of storing and exchanging information is written words, text mining has a very high commercial potential. In fact, a recent study indicated that 80% of a company's information was contained in text documents, such as emails, memos, customer correspondence, and reports. The ability to distil this untapped source of information provides substantial competitive advantages for a company to succeed in the era of a knowledge-based economy. There are many possible applications of text mining technology. We briefly highlight a few below.

1. Customer profile analysis, e.g., mining incoming emails for customers' complaint and feedback.
2. Patent analysis, e.g., analyzing patent databases for major technology players, trends, and opportunities.
3. Information dissemination, e.g., organizing and summarizing trade news and reports for personalized information services.
4. Company resource planning, e.g., mining a company's reports and correspondences for activities, status, and problems reported.

Text mining is a challenging task as it involves dealing with text data that are inherently unstructured and fuzzy. The field is interdisciplinary, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining. To facilitate discussion, this article presents a general framework for text mining consisting of two components:

Text refining: That transforms free-form text documents into an intermediate form;

Knowledge distillation: That deduces patterns or knowledge from the intermediate form.

We then use the proposed framework to study and align the state-of-the-art text mining products and applications based on the text refining and knowledge distillation functions as well as the intermediate form that they adopt. The rest of this paper is organized as follows. Section 2 presents the proposed text mining framework that bridges the gap between text mining and data mining. Section 3 gives an overview of the current text mining products and applications in the light of the proposed framework. The final section discusses open problems and research directions.

2. A TEXT MINING FRAMEWORK

Text mining can be visualized as consisting of two phases: Text refining that transforms free-form text documents into a chosen intermediate form, and knowledge distillation that deduces patterns or knowledge from the intermediate form (Tan, 1999).Intermediate form (IF) can be semi-structured such as the conceptual graph representation, or structured such as the relational data representation. Intermediate form can be document-based wherein each entity represents a document, or concept-based where in each entity represents an object or concept of interests in a specific domain. Mining a document-based IF deduces patterns and relationship across documents. Document clustering/ visualization and categorization are examples of mining from a document-based IF. Mining a concept-based IF derives pattern and relationship across objects or concepts. Data mining operations, such as predictive modeling and associative discovery, fall into this category.

A document-based IF can be transformed into a concept-based IF by realigning or extracting the relevant information according to the objects of interests in a specific domain. It follows that document-based IF could be domain-independent whereas concept-based IF is always domain-dependent.

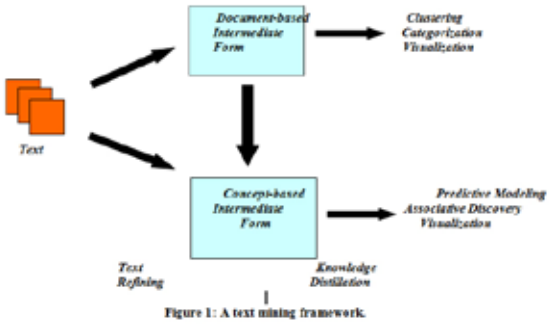

Figure 1: A text mining framework.

Figure 1: A text mining framework. Text refining converts unstructured text documents into an intermediate form (IF). IF can be document-based or

concept based. Knowledge distillation from a document-based IF deduces patterns or knowledge across documents. A document-based IF can be projected onto a concept-based IF by extracting object information relevant to a domain. Knowledge distillation from a concept-based IF deduces patterns or knowledge across objects or concepts. For example, given a set of news articles, text refining first converts each document into a document-based IF. One can then perform knowledge distillation on the document based IF for the purpose of organizing the articles, according to their content, for Visualization and navigation purposes. For knowledge discovery in a specific domain, the document-based IF of the news articles can be projected onto a concept-based IF depending on the task requirement. For example, one can extract information related to"company" from the document-based IF and form a company database. Knowledge distillation can then be performed on the company database (company-based IF) to derive Company-related knowledge.

3. INDUSTRIAL APPROCH

The text mining products/applications can be roughly organized into two groups.

1. First group focuses on document organization, visualization, and navigation.
2. Second group focuses on text analysis functions, notably, information retrieval, information extraction, categorization, and summarization.

While we see that most text mining systems provide Natural Language Processing (NLP) functions, few, if any, have integrated data mining functions for knowledge distillation across concepts or objects.

3.1. Document exploration tools

Document exploration tools organize documents based on their content and provide an environment for a user to navigate and browse in a document or concept space. A popular approach is to perform clustering on the documents based on their similarities in content and present the groups or clusters of the documents in certain graphical representation. There are a good number of text mining products that fall into this category. The following list is by no means exhaustive but should be sufficient to illustrate the variety of the representation schemes available.

Cartia's ThemeScape is an enterprise information mapping application that presents clusters of documents in a landscape representation. Canis's cMap is a document clustering and visualization tool based on Self-Organizing Map. IBM's Technology Watch, developed jointly with Synthema in Italy, is a text mining application in the scientific domain. It performs document clustering plus visualization in the form of maps for patent databases and technical publications. Insight also offers a visualization tool, known as VizControls, that performs value-added post-processing of search results by clustering the documents into groups and displaying based on a hyperbolic tree representation. Semio Corp's SemioMap employs a three-dimensional graphical interface that maps the links between concepts in the document collection. Note that SemioMap is concept-based in the sense that it explores the relationships between concepts whereas most other visualization tools are document-based.

3.2. Document analysis tools

Document analysis tools analyze the content of the documents and discover relationships between concepts or entities described in the documents. They are mainly based on natural language processing techniques, including text analysis, text categorization, information extraction, and summarization. Knowledge Discovery System's Concept Explorer is a visual search tool that helps to find precisely related content on the web. It "learns" relationships between words and phrases automatically from sample documents and visually guides you to construct searches. Insight's Linguists is another document retrieval tool with some text analysis and Summarization capabilities. IBM's Intelligent Miner is probably one of the most comprehensive text mining products around. It offers a set of text analysis tools, including a feature extraction tool, a set of clustering tools, a summarization tool, and a categorization tool. Also incorporated are the IBM's text search engine, Net Question Solution and the IBM web crawler package. TextWise, an R&D company based in Syracuse University offers various text mining products. DR-LINK is an information retrieval system based on automatic concept expansion. CINDOR is its cross lingual version. CHESS is a text analysis and information extraction tool. Also an information extraction tool is the Data Junction's Cambio, which extracts data in the form of relational attributes from text. Megaputer's Text Analyst uses a semantic net representation of documents and performs automated indexing, topic assignment, text abstraction, and semantic search.

4. OPEN PROBLEMS AND FUTURE DIRECTIONS

Despite the great potential and the mushrooming of text mining products, there are technical issues to be overcome before text mining becomes a main stream technology.

4.1. Intermediate form

Intermediate forms with varying degrees of complexity are suitable for different mining purposes. For a fine-grain domain-specific knowledge discovery task, it is necessary to perform semantic analysis to derive a sufficiently rich representation to capture the relationship between the objects or concepts described in the documents. However, semantic analysis methods are computationally expensive and often operate in the order of a few words per second. It remains a challenge to see how semantic analysis can be made much more efficient and scalable for very large text corpora.

4.2. Multilingual text refining

Whereas data mining is largely language independent, text mining involves a significant language component. Multilingual text mining is the area we expect to see a lot of activities in the next few years due to the substantial competitive advantages and the huge commercial potential that one can obtain through mining in languages other than English. Languages that are of particular interests include European languages and Asian languages, in particular Japanese and Chinese. As each language has a different syntactic structure and requires specialized semantic interpretation, a systematic approach for bringing in language modeling is inevitable and will form an essential part of multilingual text mining.

4.3. Domain knowledge integration

Current text mining systems do not make use of domain knowledge. We expect it to be an integral component of the future text mining tools. Domain knowledge is useful in orientating and focusing attention so as to improve the text parsing efficiency and to help to derive a more compact representation. Domain knowledge also plays a major role in knowledge distillation tasks. In a classification or predictive modeling task, for example, domain knowledge helps to improve learning/mining efficiency as well as the quality of the learned model (or mined knowledge) (Tan, 1997). It is also interesting to explore how a user's knowledge can be used to initialize a system's knowledge structure and make the discovered knowledge more interpretable.

4.4. Personalized autonomous mining

Another important dimension of research is to make text mining tools more user friendly. Current text mining products/applications are designed for trained knowledge specialists. Future text mining tools, as part of the knowledge management systems, should be readily usable by technical users as well as management executives. There have been some efforts in developing systems that interpret natural language queries and perform appropriate mining operations automatically. Text mining tools could also embed in intelligent personal assistants (Tan & Teo, 1998). Under the agent paradigm, a personal miner would learn a user's profile, conduct text mining operations automatically, and forward information without requiring an explicit request from the user.

CONCLUSIONS

At last we conclude that, Text mining is also known as Text Data Mining or Knowledge-Discovery in Text (KDT) [6], refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge. The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding. Most of

Text Mining operations are as follows:

- Feature Extraction.
- Text-base navigation.
- Search and Retrieval
- Categorization(Supervised Classification)
- Clustering(Unsupervised Classification)
- Summarization

Primary objective of the Feature Extraction operation is to identify facts and relations in text. Most of the times this includes distinguishing which noun phrase is a person, place, organization or other distinct object. Feature Extraction algorithms may use dictionaries to identify some terms and linguistic patterns to detect others. The Text-base navigation enables users to move about in a document collection by relating topics and that exposes selected types of semantic patterns in a document collection. To visualize text documents, we first need to convey to the analyst the underlying relation-ships between documents in a geometrically intuitive manner—we must preserve certain Visualization characteristics for the rendering to be meaningful. For example, documents that are close to each other in content should also be geometrically close to each other. Additionally, the analyst should be able to provide his or her own insight in determining what it means for documents to have close content or similar meaning. For users to provide their own insight, they have to access the meaning of the Visualization. That is, users must beable to interpret the rendered data so that the topic document relations are clearly defined.

REFERENCES

[1] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), "From data mining to knowledge discovery: An overview", in U. Fayyad et al. (eds.) Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, Mass., 1-36. | | [2] Feldman, R. & Dagan, I. (1995), "Knowledge discovery in textual databases (KDT)", in Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, AAAI Press, 112-117. | | [3] Hearst, M.A. (1997), "Text data mining: Issues, techniques, and the relationship to information access", Presentation notes for UW/MS workshop on data mining, July 1997. Simoudis, E. (1996), "Reality check for data mining", IEEE Expert, 11(5). | | [4] Tan, A.-H. (1997), "Cascade ARTMAP: Integrating neural computation and symbolic knowledge processing", IEEE Transactions on Neural Networks, 237-250. | | [5] Tan, A.-H. & Teo, C. (1998), "Learning user profiles for personalized information dissemination", in Proceedings, International Joint Conference on Neural Networks (IJCNN'98), Alaska, 183-188. | [6] Tan, A.-H. (1999), "Text Mining: The state of the art and the challenges", in Proceedings, PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, Beijing, April, 1999. | | [7] N. Kanya and S. Geetha† (2007), "Information Extraction: A Text Mining Approach", IET-UK International Conference on Information and Communication Technology in Electrical Sciences, IEEE, Dr. M.G.R. University, Chennai, Tamil Nadu, India,1111-1118. | | [8] Shantanu Godbole, and Shourya Roy, India (2008), "Text to Intelligence: Building and Deploying a Text Mining | Solution in the Services Industry for Customer Satisfaction Analysis", IEEE, 441-448. Allen, J. (1995). Natural Language Understanding, Second Edition. Redwood City, CA: Benjamin/Cummings. | [1] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), "From data mining to knowledge discovery: An overview", in U. Fayyad et al. (eds.) Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, Mass., 1-36. | | [2] Feldman, R. & Dagan, I. (1995), "Knowledge discovery in textual databases (KDT)", in Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, AAAI Press, 112-117. | | [3] Hearst, M.A. (1997), "Text data mining: Issues, techniques, and the relationship to information access", Presentation notes for UW/MS workshop on data mining, July 1997. Simoudis, E. (1996), "Reality check for data mining", IEEE Expert, 11(5). | | [4] Tan, A.-H. (1997), "Cascade ARTMAP: Integrating neural computation and symbolic knowledge processing", IEEE Transactions on Neural Networks, 237-250. | | [5] Tan, A.-H. & Teo, C. (1998), "Learning user profiles for personalized information dissemination", in Proceedings, International Joint Conference on Neural Networks (IJCNN'98), Alaska, 183-188. | [6] Tan, A.-H. (1999), "Text Mining: The state of the art and the challenges", in Proceedings, PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, Beijing, April, 1999. | | [7] N. Kanya and S. Geetha† (2007), "Information Extraction: A Text Mining Approach", IET-UK International Conference on Information and Communication Technology in Electrical Sciences, IEEE, Dr. M.G.R. University, Chennai, Tamil Nadu, India,1111-1118. | | [8] Shantanu Godbole, and Shourya Roy, India (2008), "Text to Intelligence: Building and Deploying a Text Mining | Solution in the Services Industry for Customer Satisfaction Analysis", IEEE, 441-448. Allen, J. (1995). Natural Language Understanding, Second Edition. Redwood City, CA: Benjamin/Cummings. |



Sara Publishing Academy

GLOBAL RESEARCH ANALYSIS

Multi-Subject Journal

Editor,

GLOBAL RESEARCH ANALYSIS

8-A, BANAS, Opp. SLU Girls College, Nr.Congress Bhavan, Paldi, Ahmedabad - 380006. Gujarat (INDIA)

Cell: +91 9824702127, 8866003636

Printed at Unique Offset, Novatsing Rupam Estate, Opp. Abhay Estate, Tavdipura, Shahibaug, Ahmedabad

Designed By : AdVoice Inc. +91 9824484447