A Study on Information Retrieval Methods in Text Mining

[1]Dr.M.Suresh Babu,
Professor & Head, Department of Computer Applications,
Madanapalle Institute of Technology & Science,
Madanapalle,India. .

[2]Mr. A.Althaf Ali,
Asst.Professor,Department of Computer Applications,Madanapalle Institute of Technology & Science,
Madanapalle,AP.India .

[3]Mr. A.Subramaneswara Rao,
Asst Professor,
Department of Computer Applications,
Madanapalle Institute of Technology & Science,
Madanapalle, AP. India.

***Abstract* - Information in the legal domain is often stored as text in relatively unstructured forms. For example, statutes, judgments and commentaries are typically stored as free documents. Discovering knowledge by the automatic analysis of free text is a field of research that is evolving from information retrieval and is often called text mining. Dozier states that text mining is a new field and there is still debate about what its definition should be. Jockson observe that text mining involves discovering something interesting about the relationship between the text and the world. Hearst proposes that text mining involves discovering relationships between the content of multiple texts and linking this information together to create new information. Text information retrieval and data mining has thus become increasingly important. In this paper various information retrieval techniques based on Text mining have been presented.**

Keywords: Information Retrieval, Information Extraction and Indexing Techniques

1.0 INTRODUCTION

Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web. Nowadays most of the information in government, industry, business, and other institutions are stored electronically, in the form of text databases. Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. For example, a document may contain a few structured fields, such as title, authors, publication date, and category, and so on, but also contain some largely unstructured text components, such as abstract and contents. There have been a great deal of studies on the modeling and implementation of semi structured data in recent database research. Moreover, information retrieval techniques, such as text indexing methods, have been developed to handle unstructured documents. Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user

Text mining often includes the following techniques:

a. Information extraction is a technique for extracting domain specific information from texts. Text fragments are mapped to field or template lots that have a definite semantic technique :
b. Text summarization involves identifying, summarizing and organizing related text so that users can efficiently deal with information in large documents;
c. Text categorization involves organizes documents into a taxonomy, thus allowing for more efficient searches. It involves the assignment of subject descriptors or classification codes or abstract concepts to complete texts;
d. Text clustering involves automatically clustering documents into groups where documents within each group share common features.

All text mining approaches utilize information retrieval mechanisms. Indeed, the distinction between information retrieval methods and text mining is blurred. In the next section information retrieval basics are discussed. A number of sophisticated extensions to basic information retrieval advanced in the legal field are described. We then discuss examples of information extraction, text summarization, text categorization and text clustering in law.

Information Retrieval Basics

The aim of efficient information retrieval should be to retrieve that information, and only that information which is deemed relevant to a given query. Salton states that a typical information retrieval system selects documents from a collection in response to a user's query and ranks these documents according to their relevance to the query. This is primarily accomplished by matching a text representation with a representation of the query.

Information retrieval and database systems have some similarities. Whereas database systems have focused on query processing and transactions relating to structured data, information retrieval is concerned with the organization and information from a large number of text based documents. The task of querying databases and text retrieval systems is very different. For text retrieval systems, the matching is not deterministic and often incorporates an element of uncertainty. Retrieval models generally rank the retrieved document according to their potential relevancy to the query.

Legal information retrieval considers searching both structured and unstructured content. For structured information, the semantics can be clearly and unambiguously determined and can be described with simple and clear concepts. This information category comprises, for instance, identification data of the texts, data for version management and the function and role of certain components. These data are often added in the form of metadata (i.e data that describe other data) to the documents. Unstructured information often occurs in natural language texts or in other formats such as audio and video and generally has a complex semantics. A detailed analysis of information retrieval in law can be found in Zelenznikow and Hunter and Moens.

Moens notes that the majority of existing automatic indexing methods select natural language index terms from document texts. The indexed terms selected concern single words and multi word phases and are assumed to reflect the content of the text. She claims that a prevalent process of selecting natural language index terms from texts that reflect its content is composed of the following steps:

a. Lexical analysis – the text is parsed and individual words are recognized.
b. The removal of stopwords - a text retrieval system often associates a stop list with a set of documents. A stop list is a set of words that are deemed irrelevant (such as 'a' , 'the', 'for') or at least irrelevant for the given query.
c. The optional reduction of the remaining words to their stem form – A group of different words may share the same word stem. The text retrieval system needs to identify groups of words that have a small syntactic variation from each other and only use one word from each group of. only use breach instead of breaches, breach, breached. There are different methods of stemming, many of which rely upon linguistic knowledge of the collection's language.
d. The optional formulation of phrases as index terms. Techniques of phrase recognition employ the statistics of co-occurrences of words or rely upon linguistic knowledge of the collection's language.
e. The option replacement of words, word stems or phrases by their thesaurus class terms - A thesaurus replaces the individual words or phrases of a text by more uniform concepts.
f. The computation of the importance indicator or term weight of each remaining word stem or word, thesaurus class term or phrases term.

1.1Text Databases and Information Retrieval:

Text databases (document databases)

Large collections of documents from various sources: news articles, research papers, books, digital libraries, E-mail messages, and Web pages, library database, etc.

Data stored is usually semi-structured and Traditional search techniques become inadequate for the increasingly vast amounts of text data

1.2 Information retrieval (IR)

A field developed in parallel with database systems

Information is organized into (a large number of) documents

IR deals with the problem of locating relevant documents with respect to the user input or preference

□ IR Systems and DBMS deal with different problems

Typical DBMS issues are update, transaction management, complex objects

Typical IR issues are management of unstructured documents, approximate search using keywords and relevance

□ Typical IR systems

Online library catalogs

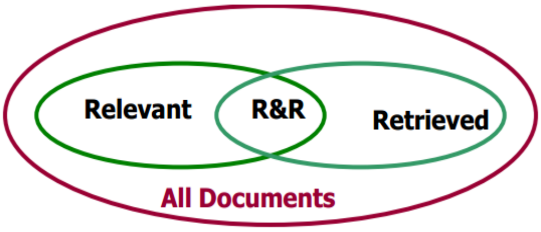
Online document management systems

□ Main IR approaches

"pull" for short-term information need

"push" for long-term information need (e.g., recommender systems)

Basic methods of Text Retrieval :



$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

$$F_{score} = \frac{recall \times precision}{(recall + precision)/2}$$

1.3 Text Retrieval Methods

□ Document Selection (keyword-based retrieval)
Query defines a set of requisites
Only the documents that satisfy the query are returned
A typical approach is the Boolean Retrieval Model
□ Document Ranking (similarity-based retrieval)
Documents are ranked on the basis of their relevance with respect to the user query
For each document a "degree of relevance" to the query is measured
A typical approach is the Vector Space Model.

1.4 Boolean Retrieval Model

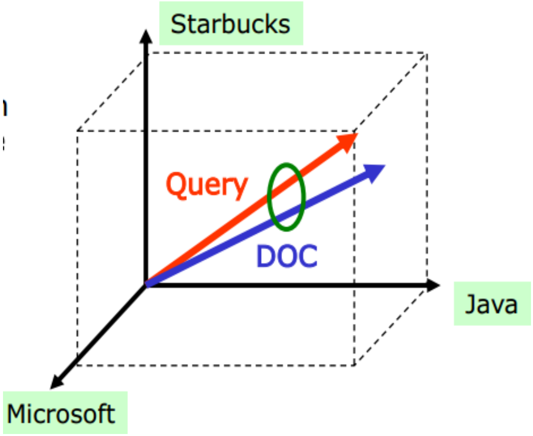
□ A query is composed of keywords linked by the three logical connectives: not, and, or
E.g.: "car and repair", "plane or airplane"
□ In the Boolean model each document is either relevant or non-relevant, depending it matches or not the query
□ Limitations
Generally not suitable to satisfy information need Useful only in very specific domain where users have a big expertise

1.5 Vector Space Model :

A document and a query are represented as vectors in high dimensional space corresponding to all the keywords. Relevance is measured with an appropriate similarity measure defined over the vector space.

Issues :

How to select keywords to capture "Basic concepts"?
How to assign weights to each term?
How to measure the similarity?



2.0 Keywords Selection

□ Text is preprocessed through tokenization
□ Stop list and word stemming are used to identify significant keywords
Stop List
• e.g. "a", "the", "always", "along"
Word stemming
• e.g. "computer", "computing", "computerize" => "compute"

Keywords Weighting
□ Term Frequency (TF)
Computed as the frequency of a term t in a document d (or as the relative frequency)
More frequent a term is □ more relevant it is
□ Inverse Document Frequency (IDF)

$$\mathrm{IDF}(t) = \log \frac{|D|}{1 + |D_t|}$$

D is the documents collection, Dt is the subset of D that contains t
Less frequent among documents □ more discriminant
• e.g., database in a collection of papers on DBMS
□ Mixing TF and IDF

$$\text{TF-IDF}(d,t) = \mathrm{TF}(d,t) \times \mathrm{IDF}(t)$$

How to measure similarity :
Given two documents (or a document and a query)

$$D_i = (w_{i1}, w_{i2}, \cdots, w_{iN}) \qquad D_j = (w_{j1}, w_{j2}, \cdots, w_{jN})$$

Similarity Definition
Dot product

$$Sim(D_i, D_j) = \sum_{t=i}^{N} w_{it} * w_{jt}$$

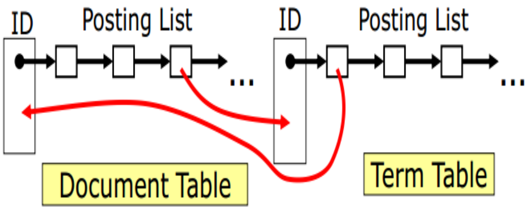
normalized dot product (or cosine)

$$Sim(D_i, D_j) = \frac{\sum_{t=i}^{N} w_{it} * w_{jt}}{\sqrt{\sum_{t=1}^{N} (w_{it})^2 * \sum_{t=1}^{N} (w_{jt})^2}}$$

2.1 Text Indexing

□ Inverted index

Maintains 2 tables:



Implemented with hash tables or B+ trees

• Find all docs associated to one or to a set of terms

• Find all terms associated to a doc

2.2 Signature file

Signature bitstring for each document

Each bit represents one or terms (1 if present 0 otherwise) and Signatures are used to retrieve an initial match of the query.

2.3 Dimensionality Reduction

Approaches presented so far involves high dimensional space (huge number of keywords)

Computationally expensive and Difficult to deal with synonymy and polysemy problems

• "vehicle" is similar to "car"

• "mining" has different meanings in different contexts

Dimensionality reduction techniques

Latent Semantic Indexing (LSI)

Locality Preserving Indexing (LPI)

Probabilistic Semantic Indexing (PLSI)

2.4 Latent Semantic Indexing (LSI)

□ Let xi be vectors representing documents and X (term frequency matrix) the all set of documents:

$$\vec{x}_1, \cdots, \vec{x}_n \in R^m \qquad X = [\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n]$$

□ Let use the singular value decomposition (SVD) to reduce the size of frequency table: $X = U \Sigma V^T$

Approximate X with Xk that is obtained from the first K vectors of U

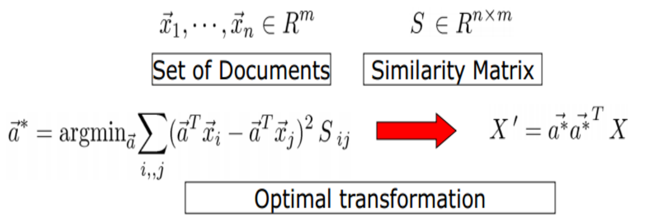
It can be shown that such transformation minimizes the error for the reconstruction of X.

2.5 Locality preserving Indexing (LPI)

□ Goal is preserving the locality information

Two documents close in the original space should be close also in the transformed space

More formally



$$\vec{x}_1, \cdots, \vec{x}_n \in R^m \qquad S \in R^{n \times m}$$

$$\vec{a}^* = \operatorname{argmin}_{\vec{a}} \sum_{i,,j} (\vec{a}^T \vec{x}_i - \vec{a}^T \vec{x}_j)^2 S_{ij} \qquad X' = \vec{a}^* \vec{a}^{*T} X$$

Similarity Matrix :

$$S_{ij} = \begin{cases} \frac{\vec{x}_i^T \vec{x}_j}{||\vec{x}_i^T \vec{x}_j||}, \\ 0, \end{cases}$$

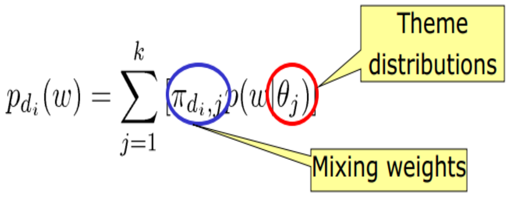
if $\vec{x}_i$ is in the p nearest neighbors of $\vec{x}_j$ or viceversa otherwise.

Probabilistic Latent Semantic Indexing (PLSI)

□ Similar to LSI but does not apply SVD to identify the k most relevant features

□ Assumption: all the documents have k common "themes"

□ Word distribution in documents can be modeled as



$$p_{d_i}(w) = \sum_{j=1}^{k} [\pi_{d_i,j} p(w|\theta_j)]$$

Mixing weights are identified with Expectation-Maximization (EM) algorithms and define new representation of the documents

Usage of Text Mining :

Text Mining aims to extract useful knowledge from text documents

□ Approaches

Keyword-based

• Relies on IR techniques

Tagging

• Manual tagging

• Automatic categorization

Information-extraction

• Natural Language Processing (NLP)

Keyword-Based Association Analysis

Document Classification

Natural Language Processing is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

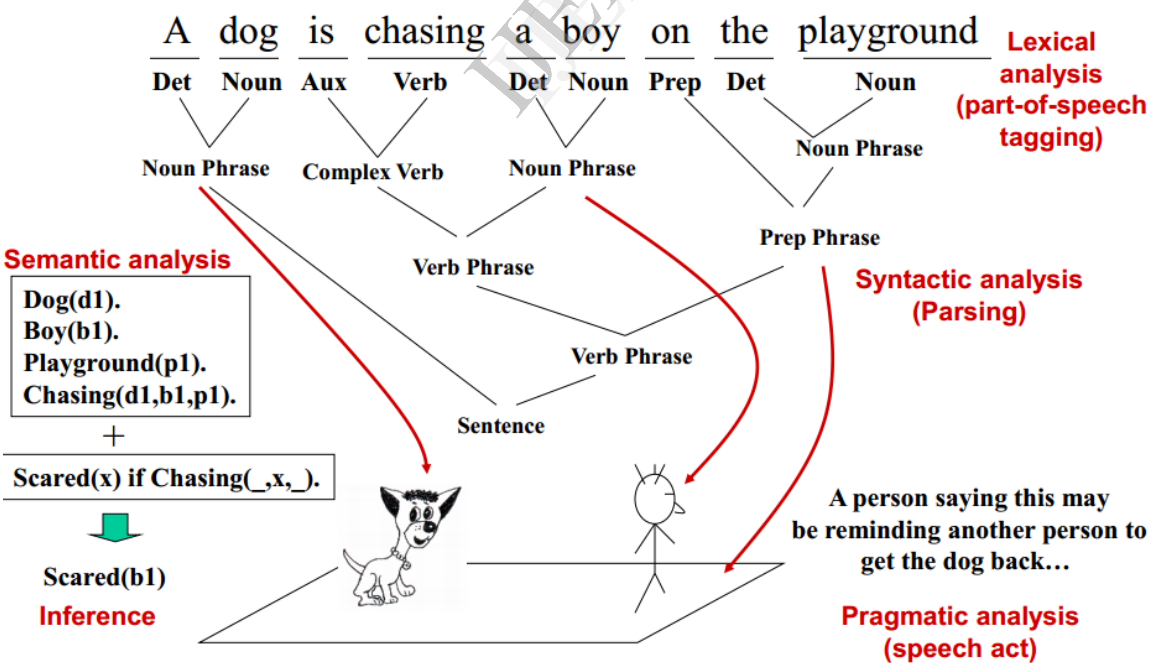
Modern NLP algorithms are based on machine learning, especially statistical machine learning. The paradigm of machine learning is different from that of most prior attempts at language processing. Prior implementations of language-processing tasks typically involved the direct hand coding of large sets of rules. The machine-learning paradigm calls instead for using general learning algorithms — often, although not always, grounded in statistical inference — to automatically learn such rules through the analysis of large *corpora* of typical real-world examples. A *corpus* (plural, "corpora") is a set of documents (or sometimes, individual sentences) that have been hand-annotated with the correct values to be learned.

Many different classes of machine learning algorithms have been applied to NLP tasks. These algorithms take as input a large set of "features" that are generated from the input data.

Some of the earliest-used algorithms, such as decision trees, produced systems of hard if-then rules similar to the systems of hand-written rules that were then common. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage that they can express the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system.

Systems based on machine-learning algorithms have many advantages over hand-produced rules:

- The learning procedures used during machine learning automatically focus on the most common cases, whereas when writing rules by hand it is often not obvious at all where the effort should be directed.
- Automatic learning procedures can make use of statistical inference algorithms to produce models that are robust to unfamiliar input (e.g. containing words or structures that have not been seen before) and to erroneous input (e.g. with misspelled words or words accidentally omitted). Generally, handling such input gracefully with hand-written rules — or more generally, creating systems of hand-written rules that make soft decisions — is extremely difficult, error-prone and time-consuming.
- Systems based on automatically learning the rules can be made more accurate simply by supplying more input data. However, systems based on hand-written rules can only be made more accurate by increasing the complexity of the rules, which is a much more difficult task. In particular, there is a limit to the complexity of systems based on hand-crafted rules, beyond which the systems become more and more unmanageable. However, creating more data to input to machine-learning systems simply requires a corresponding increase in the number of man-hours worked, generally without significant increases in the complexity of the annotation process.


A dog is chasing a boy on the playground
Det Noun Aux Verb Det Noun Prep Det Noun
Lexical analysis (part-of-speech tagging)
Noun Phrase
Complex Verb
Noun Phrase
Noun Phrase
Prep Phrase
Verb Phrase
Verb Phrase
Sentence
Semantic analysis
Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).
+
Scared(x) if Chasing(_,x,_).
Scared(b1)
Inference
Syntactic analysis (Parsing)
A person saying this may be reminding another person to get the dog back…
Pragmatic analysis (speech act)

3.0 Obstacles to NLP
Ambiguity

A man saw a boy with a telescope.
□ Computational Intensity
Imposes a context horizon.
□ Text Mining NLP Approach

Locate promising fragments using fast IR methods (bag-of-tokens)
Only apply slow NLP techniques to promising fragments.

3.1Keyword-Based Association Analysis :

Aims to discover sets of keywords that occur frequently together in the documents and Relies on the usual techniques for mining associative and correlation rules

Each document is considered as a transaction of type {document id, {set of keywords}}

Association mining may discover set of consecutive or closely-located keywords, called terms or phrase

Compound (e.g., {Stanford,University})

Noncompound (e.g., {dollars,shares,exchange})

Once discovered the most frequent terms, term-level mining can be applied most effectively (w.r.t. single word level).

3.2 Document classification

Solve the problem of labeling automatically text documents on the basis of

- Topic
- Style
- Purpose

Usual classification techniques can be used to learn from a training set of manually labeled documents

Which features? Keywords can be thousands…

□ Major approaches

- Similarity-based
- Dimensionality reduction
- Naïve Bayes text classifiers.

3.3 Similarity-based Text Classifiers

Exploits IR and k-nearest-neighbor classifier

For a new document to classify, the k most similar documents in the training set are retrieved and Document is classified on the basis of the class distribution among the k documents retrieved

• Majority vote

• Weighted vote

Tuning k is very important to achieve a good performance

The Limitations of Similarity based text classifiers are: Space overhead to store all the documents in training set and Time overhead to retrieve the similar documents.

3.4 Dimensionality Reduction for Text

Classification

□ As in the Vector Space Model used for IR, the goal is to reduce the number of features to represents text

□ Usual dimensionality reduction approaches in IR are based on the distribution of keywords among the whole documents database

□ In text classification it is important to consider also the correlation between keywords and classes and Rare keywords have an high TF-IDF but might be uniformly distributed among classes

LSI and LPI do not take into account classes distributions

□ Usual classification techniques can be then applied on reduced features space:

SVM

Bayesian classifiers

Naïve Bayes for Text

□ Definitions

Category Hypothesis Space: H = {C1, …, Cn}

Document to Classify: D

Probabilistic model:

$$P(C_i \mid D) = \frac{P(D \mid C_i)P(C_i)}{P(D)}$$

We choose the class C* such that

$$C^* = \arg\max_C P(C|D) = \arg\max_C P(D|C)P(C)$$

Issues

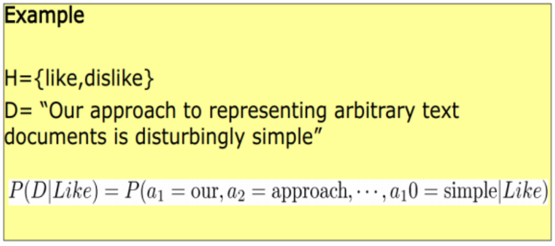
Which features?

How to compute the probabilities?

Features can be simply defined as the words in the document

□ Let ai be a keyword in the doc, and wj a word in the vocabulary, we get:

$$P(D|C) = P(a_1 = w_{j_1}, a_2 = w_{j_2}, \cdots, a_n = w_{j_n}|$$



Example

H={like,dislike}

D= “Our approach to representing arbitrary text documents is disturbingly simple”

$P(D|Like) = P(a_1 = \text{our}, a_2 = \text{approach}, \cdots, a_10 = \text{simple}|Like)$

Features can be simply defined as the words in the document

□ Let ai be a keyword in the doc, and wj a word in the vocabulary, we get:

$$P(D|C) = P(a_1 = w_{j_1}, a_2 = w_{j_2}, \cdots, a_n = w_{j_n}|C)$$

□ Assumptions

Keywords distributions are inter-independent

Keywords distributions are order-independent

$$P(D|C) = \prod_{i=1}^{n} P(w_{j_i}|C)$$

Simply counting the occurrences may lead to wrong results when probabilities are small

□ M-estimate approach adapted for text:

$$P(w_k|C) = \frac{N_{c,k} + 1}{N_c + |\text{Vocabulary}|}$$

Nc is the whole number of word positions in documents of class C, Nc,k is the number of occurrences of wk in documents of class C and |Vocabulary| is the number of distinct words in training set here Uniform priors are assumed

Final classification is performed

$$C^* = \arg\max_C P(C) \prod_{i=1}^{n} P(w_{j_i}|C)$$

Despite its simplicity Naïve Bayes classifiers works very well in practice for Newsgroup post classification and NewsWeeder (news recommender).

4.0 CONCLUSION

Most of knowledge hidden in electronic media of an organization is encapsulated in documents. Acquiring this knowledge implies effective querying of the documents as well as the combination of information pieces from different textual sources (e.g.: the World Wide Web). Discovering such hidden know ledge is an essential requirement for many corporations, due to its wide spectrum of applications. In this paper, the notion of text mining and several information retrieval techniques have been introduced and several techniques available have been presented. Due to its novelty, there are many potential research areas in the field of Text Mining, which includes finding better intermediate forms for representing the outputs of information extraction, an XML document may be a good choice. Mining texts in different languages is a major problem, since text mining tools should be able to work with many languages and multilingual documents. Integrating a domain knowledge base with a text mining engine would boost its efficiency, especially in the information retrieval and information extraction phases.

5.0 REFERENCES

[1] M. A. Hearst. What is text mining? http://www.sims.berkeley.edu/˜hearst/text-mining.html, Oct. 2003.

[2] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann Publishers, San Francisco, CA, 1999.

[3] R.Baeza-Yates and B.Ribeiro-Neto, Modern Information Retrieval addition-Wesley,Boston,1999

[4] Jochen Dorre, Peter Gersti, Roland Seiffert (1999), Text Mining: Finding Nuggets in Mountains of Textual Data, ACM KDD 1999 in San Diego, CA, USA.

[5] Ah-Hwee Tan, (1999), Text Mining: The state of art and the challenges, In proceedings, PAKDD'99 Workshop on Knowledge discovery from Advanced Databases (KDAD'99), Beijing, pp. 71-76, April 1999.

[6] Danial Tkach, (1998), Text Mining Technology Turning Information Into Knowledge A white paper from IBM .

[7]. Helena Ahonen, Oskari Heinonen, Mika Klemettinen, A. Inkeri Verkamo, (1997), Applying Data Mining Techniques in Text Analysis, Report C-1997-23, Department of Computer Science, University of Helsinki, 1997

[8]. Mark Dixon,(1997), An Overview of Document Mining Technology, http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dixm

[9] Juan José García Adeva and Rafael Calvo, "Mining Text with Pimiento", University of Sydney

[10] Text Mining Application Programming by Manu Konchadi. Published by Charles River Media. ISBN: 1584504609

[11] Jeong, D.H, Hwang, M., Kim, J., Song, S.K., Jung, H., Peters, C., Pietras, N., Kim, D.W.: Information Service Quality Evaluation Model from the User's Perspective, The 2nd International Semantic Technology (JIST) Conference 2012, Nara, Japan, 2012.

[12] ISO 9126:1-3 : Software Quality Model.

[13] ISO 25000 : Software Quality Requirement and Evaluation.

[14] ISO 14598 : Information Technology - Software Product Evaluation.

[15] ISO 9000 : Quality Management.

[16] ISO 14001 : Environmental Management.

[17] Quality of Service (W3C), Home page at, http://www.w3.org/Architecture/qos.html, viewed September 25 2012.

[18] Saroja, G., Sujatha, G.: Application of total quality management to library and information services in Indian open universities, http://www.col.org/forum/pcfpapers/saroja.pdf, viewed July 15 2012.

[19] Caruana, A.: Service loyalty: The effects of service quality and the mediating role of customer