Research Journal of Applied Sciences, Engineering and Technology 10(10): 1164-1174, 2015
DOI: 10.19026/rjaset.10.1884
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: February 16, 2015 Accepted: March 12, 2015 Published: August 05, 2015

Research Article

A Survey on Web Text Information Retrieval in Text Mining

[1]Tapaswini Nayak, [2]Srinivash Prasad and [3]Manas Ranjan Senapati
[1]Department of Computer Science, Centurion University, India
[2]GMR Institute of Technology, India
[3]Centurion University, India

Abstract: In this study we have analyzed different techniques for information retrieval in text mining. The aim of the study is to identify web text information retrieval. Text mining almost alike to analytics, which is a process of deriving high quality information from text. High quality information is typically derived in the course of the devising of patterns and trends through means such as statistical pattern learning. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, creation of coarse taxonomies, sentiment analysis, document summarization and entity relation modeling. It is used to mine hidden information from not-structured or semi-structured data. This feature is necessary because a large amount of the Web information is semi-structured due to the nested structure of HTML code, is linked and is redundant. Web content categorization with a content database is the most important tool to the efficient use of search engines. A customer requesting information on a particular subject or item would otherwise have to search through hundred of results to find the most relevant information to his query. Hundreds of results through use of mining text are reduced by this step. This eliminates the aggravation and improves the navigation of information on the Web.

Keywords: Information retrieval, text mining, web mining, web search engine

INTRODUCTION

Text mining, which is referred to as "text analytics" is one way to make qualitative or "unstructured" data vulnerable by a computer (Vasumathi and Moorthi, 2012). Qualitative data is explanatory data that cannot be measured in numbers and often includes qualities of appearance like color, texture and textual report. Quantitative data is numerical, structured data that can be deliberate. However, there is frequently slippage between qualitative and quantitative categories (Preethi, 2014). For example, a photograph might conventionally be considered "qualitative data" but when you break it down to the level of pixels, which can be measured. An estimated 70-80% of data is unstructured? This includes customer care web chats, e-mails, mobile application or web articles, news sites, social sites, internal reports, call center logs, journal papers, blog entries, to name a few. Thanks to the web and social media, More than a trillion web pages of text are being added to our communal storehouse, daily.

The Oxford English Dictionary defines text mining as the process or practice of examining large collections of written resources in order to generate new Information, classically using specialized computer software. It is a subset of the superior field of data mining. Guernsey explains that "to the unskilled, it may seem that Google and other Web search engines do something similar, since they also minute opening from beginning to end reams of documents in split-second intervals (Wang *et al*., 2011) (Fig. 1).

They do not recommend connections or generate any new knowledge. Preethi (2014) Text-mining programs go advance, categorizing information, making links between otherwise unrelated documents and providing visual maps. Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining.

Information Retrieval (IR) systems: Recognize the documents in a collection which match a user's query. The most notorious IR systems are search engines such as Google™, which identify those documents on the WWW that are relevant to a set of given words. IR systems are frequently used in libraries, where the documents are typically not the articles themselves but digital records containing information about the articles (Sagayam *et al*., 2012). This is conversely changing with the initiation of digital libraries, where the documents being retrieved are digital versions of articles and journals. IR systems allow us to taper down the set of documents that are related to a particular problem. Though text mining involves applying very computationally exhaustive algorithms to large document collections, IR can speed up the analysis

Corresponding Author: Tapaswini Nayak, Department of Computer Science, Centurion University, India

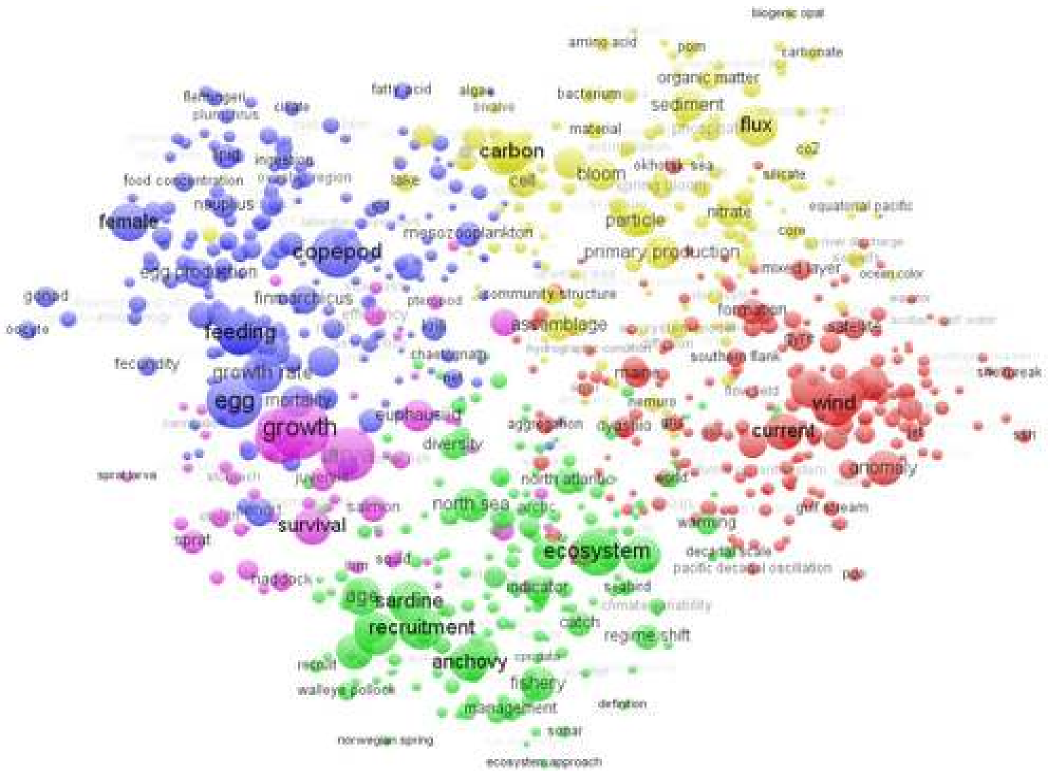

Fig. 1: Image from researchtrends.com to describe text mining

significantly by plummeting the number of documents for analysis. For example, if we are interested in mining information only about OJEE counseling, we might restrict our analysis to documents that contain the name of a OJEE, or some form of the verb 'to interact' or one of its synonyms.

This study of web text information retrieval we have analyzed different process of information retrieval in text mining (Sagayam *et al*., 2012). How the information is extracted from unstructured data to represent it in organized manner. We have also proposed potential domain for future research.

A SURVEY OF TEXT INFORMATION RETRIEVAL

The text Information Retrieval products and applications based on the text refining and knowledge refinement functions as well as the middle form adopted. One group of products focuses on document grouping, image and map-reading.

An additional group focuses on text analysis function, information retrieval, information extraction, categorization and summarization. While we see that most text Information Retrieval systems are based on usual words processing none of the products has integrated data Information Retrieval functions for information distillation across theory or objects.

IR tools on the web: Information from web can be retrieved by number of tools available ranging from general purpose search engines to dedicated search engines. Following are the most commonly used web IR tools.

General-purpose search engine: They are the most commonly used tool for information retrieval. Google, AltaVista, Excite are some of the examples. Each of them has its own set of web pages which they search to answer a query.

Hierarchical directories: In this approach the user is required to choose one of a given set of categories at each level to get to the next level. For example, Yahoo! or the dmoz open directory project.

Specialized search engines: These search engines are specialized on an area and provides huge collection of documents related to that specific area. For e.g., PubMed, a search engine specialized on medical publications.

Other search paradigms: There are various other search paradigms. A Search-by-Example feature exists in various incarnations. Also various collaborative filtering approaches and notification systems exist on the Web.

Use of search engine in information retrieval: Information Retrieval on the Web is a variant of

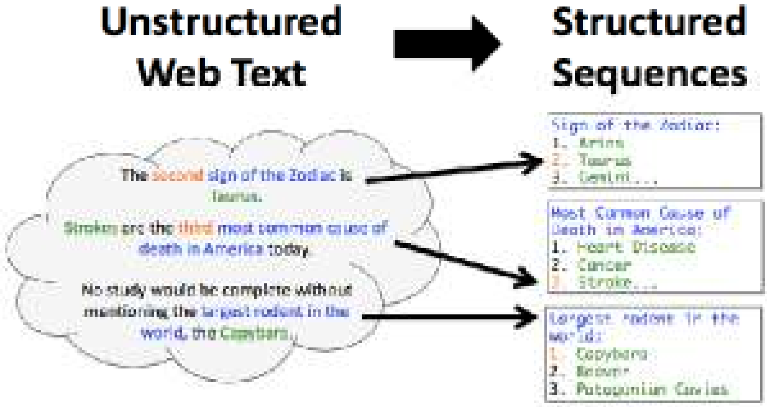
Unstructured Web Text
Structured Sequences
The second sign of the Zodiac is Taurus.
Strokes are the third most common cause of death in America today.
No study would be complete without mentioning the largest rodent in the world, the Capybara.
Sign of the Zodiac:
1. Aries
2. Taurus
3. Gemini...
Most Common Cause of Death in America:
1. Heart Disease
2. Cancer
3. Stroke...
Largest rodent in the world:
1. Capybara
2. Beaver
3. Patagonian Cavies

Fig. 2: Retrieval of information from unstructured web text

classical information retrieval. As in classical information retrieval, a user tries to satisfy an information need in a collection of documents. In this case the collection of documents consists of all the web pages in the publicly accessible web. Given a user query the goal is to retrieve high quality web pages that are relevant to the user's need. So, finding high quality documents is an additional requirement that arises in the web context.

Search engines are used to index a sizeable portion of the web across all topics and domains to retrieve the information (Barathi and Valli, 2011). Each such Engine consists of three major components: A spider or crawler (Wikipedia) browses the web by starting with a list of URLs called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs which are visited recursively to form a huge collection of documents called corpus. The corpus is typically augmented with pages obtained from direct submissions to search engines and various other sources. Each crawler has different policies with respect to which links are followed, how deep various sites are explored, etc. As a result, there is astoundingly little correlation among corpora of various engines. Figure 2 explains about the information retrieval system in search engine. The indexer processes the data and represents it usually in the form of fully upturned files.

However, each major Search Engine uses different representation schemes and has different policies with respect to which words are indexed. The query processor which processes the input query and returns matching answers, in an order determined by a ranking algorithm. It consists of a front end that transforms the input and brings it to a standard format and a back end that finds the matching documents and ranks them. Search engines allow you to direct your search to all the public materials available on the Internet via your Internet browsers. This includes: websites, Internet discussion groups and news groups, information in different formats (images, sound and video). You can limit your search by e.g., time, geography, language and/or, the so-called domain name country codes.

INFORMATION RETRIEVAL

Consequently the information retrieval system has to deal with the following tasks.

Generating structured representations of information items: This process is called feature extraction and can include simple tasks, such as extracting words from a text as well as complex system, e.g., for image or video analysis. Generating structured representations of information needs: often these tasks are solved by providing users with a query language and leave the formulation of structured queries to them. This is the case for example of simple keyword based query languages as used in Web search engine. Some information retrieval systems also maintain the user in the query formulation, e.g., through image interface. Similar of information needs with information substance: this is the algorithmic duty of computing parallel of information items and information need of constitutes the heart of the information retrieval system. Similarity of the structured representations is used to system relevance of information for users.

As a result a selection of relevant information items or a ranked result can be presented to the user. Since information retrieval systems deal usually with large information collections and large user communities the efficiency of an information retrieval system is crucial. This imposes fundamental constraints on the retrieval system. Retrieval systems that would capture relevance very well, but are computationally prohibitively expensive are not appropriate for an information retrieval system.

Text information retrieval: At present most popular information retrieval methods are Web search engine. To a huge amount, they are text retrieval system, since they exploit only the textual content of Web documents for retrieval (Ai *et al*., 2005). However, more recently Web search engines also create to exploit link information and even image information.

The three tasks of a Web search engine for retrieval are:

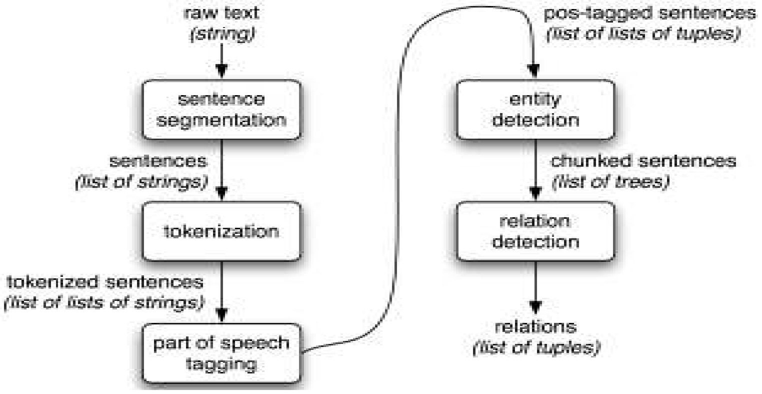

Fig. 3: Extraction of relation from raw text

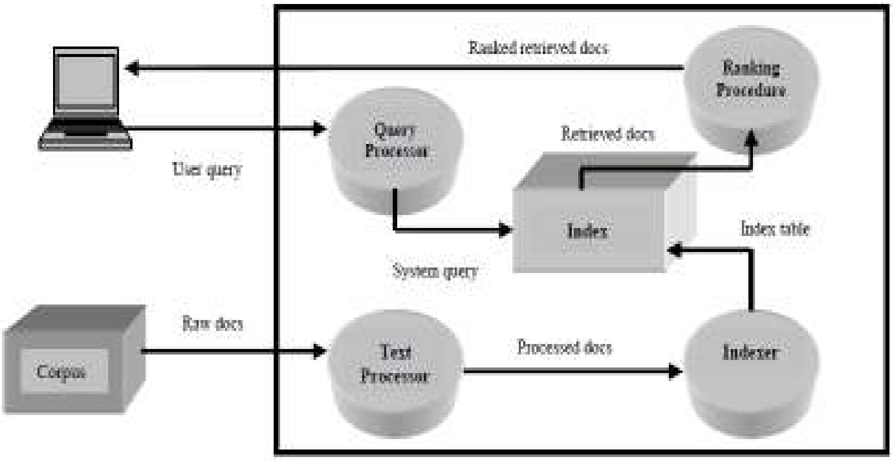

Fig. 4: Information retrieval system in search engine

- Extracting the textual features, which are the words or terms that occur in the document. We guess that the web search engine has already composed the documents from the Web using a Web crawler.
- Support the formulation of textual queries. This is usually done by allowing the entry of keywords through Web forms.
- Computing the similarity of documents with the query and producing from that a grade result. Here Web search engines are used for standard text a retrieval method that is such as Boolean retrieval and vector space retrieval. We will introduce these methods in detail subsequently.

Information extraction: The general purpose of Knowledge Discovery is to "extract implicit, previously unknown and potentially useful information from data". Information Extraction IE mainly deals with identifying words or feature terms from within a textual file. Feature terms can be defined as those which are directly related to the domain.

Comparison with conventional information retrieval: In essence the differences between conventional information retrieval system and web based information retrieval system can be partitioned into two parts, namely differences in the documents and differences in the users (Fig. 3 and 4).

Differences in the documents:
Hypertext: Documents present on the web are different from general text-only documents because of the presence of hyperlinks. It is estimated that there are at least one hyperlinks present per document.

Duplication: On the web, over 40% of the documents nearby are either near or exact duplicates of other documents and this estimation has not included the semantic duplicates yet.

Number of documents: The size of web has grown exponentially over the past few years. The collection of documents is over trillions and this collection is much larger than any collection of documents processed by an information retrieval system. According to estimation, web at present grows by 18% per month.

Heterogeneity of document: The contents present on a web page are heterogeneous in nature i.e., in addition to

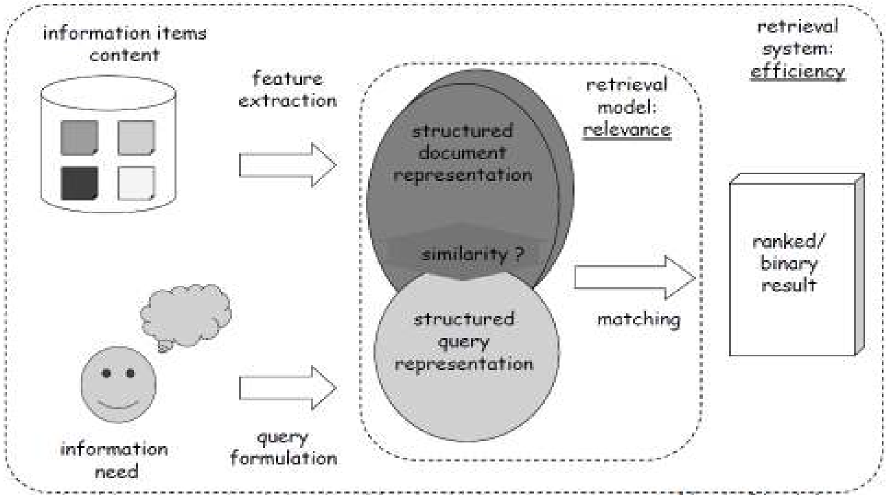
information items content
feature extraction
structured document representation
retrieval model: relevance
retrieval system: efficiency
similarity ?
matching
ranked/ binary result
structured query representation
information need
query formulation

Fig. 5: Information retrieval

text they might contain other multimedia contents like audio, video and images.

Lack of stability: Web pages lack stability means the contents of Web pages are modified frequently. Besides that any person using internet can create a web pages even if it contains authentic information or not. The users on the web act upon differently than the users of the conventional information retrieval systems. The users of the latter are mostly skilled librarians whereas the range of Web users varies from a layman to a technically sound person. Typical user activities shows.

Poor queries: Most of the queries submitted by users are generally short and be deficient in helpful keywords that may help in the retrieval of apposite information.

Reaction to results: Usually users don't estimate all the result screens, they confine to only results displayed in the first result screen.

Heterogeneity of users: There is a wide variance in web users and their web familiarity.

To meet these confronts, any type of information retrieval system either a conventional one or web based has to undergo four essential steps.

Document processing: The text present in the corpus is processed into a predefined format; stems of the words are extracted i.e., words in the document are assembled to their root words to make index entries.

Query processing: User queries are tokenized into understandable segments, these segments are parsed and a general query representation is made which is then used for matching query terms with the inverted index entries.

A search and matching function: Each document in the corpus is searched for the query terms and based on the matching of terms; each document is given some matching score. However, different systems may adopt different models for performing this searching and matching.

A ranking capability: Based on the match score of each document they are retrieved as a result of user query in the decreasing order of their relevance i.e., the documents with higher relevance will be displayed first than others with lesser relevance.

Measures for text retrieval: To evaluate the performance of our text retrieval system we use standard measures such as recall, precision and F-score measure. Let the set of documents relevant to a query be denoted as Relevant and the set of document retrieved be denoted as Retrieved. The set of documents that are both relevant and retrieved is denoted as Relevant T Retrieved, There are two basic measures for assessing the quality of text retrieval as shown in the pie-chart of Fig. 5.

Precision is the percentage of retrieved documents that are in fact relevant to the query:

$$\text{Precision} = |\{\text{Relevant}\} \cap \{\text{Retrieved}\} |/|\{\text{Retrieved}\}\text{which is always} < = 0 \quad (1)$$

Recall is the percentage of documents that are relevant to the query and were, in fact, retrieved:

$$\text{Recall} = |\{\text{Relevant}\} \cap \{\text{Retrieved}\} |/| \{\text{Relevant}\}|\text{is always } 0 \quad (2)$$

An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used tradeoff is the F-score:

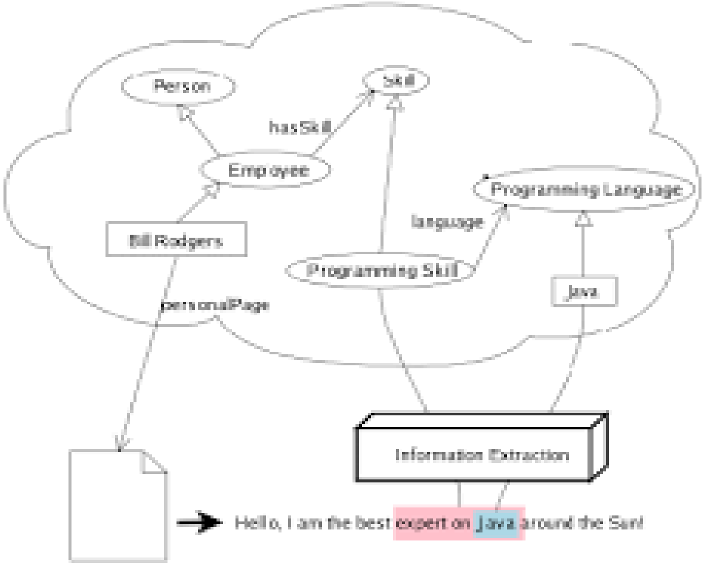

Fig. 6: Text information retrieval

$$\text{F-score} = (2\times\text{Recall}\times\text{Precision})/(\text{Recall} + \text{precision}) \quad (3)$$

For example the user's query is OJEE counseling. The set of documents relevant to the users query is 7 and the set of documents retrieved for that user's query is 37. So the Precision is {(7) ∩ (37)} /37 = 0.189 which is always <0. The Recall is {(7) ∩ (37)} /7 = 1 (always 1). The harmonic mean of precision and Recall is 1×0.189/ (1 + 0.189) /2 = 0.3179. Similarly the F-score for the query named CUTM is 0.672. These three measures are not directly useful to judge against two ranked lists of documents because they are not susceptible to the interior ranking of the documents in a retrieved set. In order to measure the quality of a ranked list of documents, it is common to compute an average of precisions at all the ranks where a new relevant document is returned. It is also common to plot a graph of precisions at many different levels of recall; a higher curve represents a better-quality information retrieval system. Figure 6 depicts the curve of Precision and recall.

KEYWORD BASED TEXT RETRIEVAL INDEXING TECHNIQUES

Most information retrieval systems support keyword based and similarity based retrieval. In keyword based information retrieval a document is represented by a string, which can be identified by asset of keywords. A user provides a keyword or an expression formed out of a set of keywords such as "STUDENTS and THEIR BOOKS". A good information retrieval system needs to consider synonyms when answering such query. This is a simple model that encounters two difficulties:

- The synonyms problem, keywords may not appear in the document, even though the document is closely related to the keywords.
- **The polysemy problem:** The same keyword may mean different things in different contexts.

The information retrieval system based on similarity finds similar documents based on a set of common keywords. The output for this system is based on the degree of relevance measured based on using keywords closeness and the relative frequency of the keywords. In some cases it is difficult to give a precise measure of the relevance between keyword sets.

One of the measure of document similarity is $Sim(v_1, v_2) = (v_1.v_2) / (||v_1||v_2||)$.
where, $v.v_2$ is the standard dot products. $||v_1|| = \sqrt{v_1.v_2}$.

To evade indexing useless words, a text retrieval system often acquaintances a stop list with a set of documents. A stop list is a set of words that are deemed extraneous. For example, the, with, for, to, a and so on are stop words, even though they may become visible recurrently. Stop lists may differ per document set. For example database systems could be a significant keyword in a news paper. However it may be considered as a stop word in a set of research paper presented in a data base conference.

Some indexing techniques such as inverted indices and signature files are used for text indexing. Where document table consists of a set of document records, each containing two fields: doc id and posting list, where posting list is a list o f terms (or pointers to terms) that occur in the document, sorted according to some relevance measure. Term table consists of a set of term records each containing two field s: doc id and posting list, where posting list is a list of terms (or pointers to terms) that occur in the document, sorted according to some relevance measure. Term table consists of a set of term records, each containing two fields: term id and posting list, where posting list specifies a list of document identifiers in which the term appears. With such organization, it is easy to answer queries like "Find all of the documents associated with a given set of terms," or "Find all of the terms associated with a given set of documents." For example, to find all of the documents associated with a set of terms, we can first find a list of document identifiers in term table for each term and then intersect them to obtain the set of relevant documents.

The posting lists could be rather long, making the storage requirement quite large. They are easy to implement, but are not satisfactory at handling synonymy (where two very different words can have the same meaning) and polysemy (where an individual word may have many meanings). A signature file is a file that provisions a signature record for each document in the database. Each signature has a fixed size of *b* bits representing terms. A simple encoding scheme goes as follows. Each bit of a document signature is initialized to 0. A bit is set to 1 if the term it represents appears in the document. A signature *S*1 matches another signature *S*2 if each bit that is set in signature *S*2 is also set in *S*1. Since there are usually more terms than available bits, multiple terms may be mapped into the same bit. Such multiple-to one mapping make the search expensive because a document that matches the signature of a query does not necessarily contain the set of keywords of the query. The document has to be retrieved, parsed, stemmed and checked. Improvements can be made by first performing frequency analysis, stemming and by filtering stop words and then using a hashing technique and superimposed coding technique to encode the list of terms into bit representation. Nevertheless, the problem of multiple-to-one mappings still exists, which is the major disadvantage of this approach.

Query processing technique: Once an inverted index is formed for a document collection, a retrieval system can response a keyword query rapidly by looking up which documents contain the query keywords. Specifically, we will maintain a score accumulator for each document and update these accumulators as we go through each query term. For each query term, we will fetch all of the documents that match the term and increase their scores. When we do not have such relevant examples, a system can assume the top few retrieved documents in some initial retrieval results to be relevant and extract more related keywords to expand a query. Such feedback is called pseudo-feedback or blind feedback and is essentially a process of mining useful keywords from the top retrieved documents. Pseudo-feedback also often leads to improved retrieval performance. One major limitation of many existing retrieval methods is that they are based on exact keyword matching. However, due to the complexity of Natural languages, keyword based retrieval can encounter two major difficulties. The first is the synonymy problem: two words with identical or similar meanings may have very different surface forms. For example, a user's query may use the word "automobile," but a relevant document may use "vehicle" instead of "automobile." The second is the polysemy problem: the same keyword, such as mining, or Java, may mean different things in different contexts.

Categorization: Categorization automatically assigns one or more category to free text document. Categorization is supervised learning method because it is based on input output examples to classify new documents. Predefined classes are assigned to the text documents based on their content. The typical text categorization process consists of pre-processing, indexing, dimensionally reduction and classification. The goal of categorization is to train classifier on the basis of known examples and then unknown examples are categorized automatically. Statistical classification techniques like Naïve Bayesian classifier, Nearest Neighbor classifier, Decision Tree and Support Vector Machines can be used to categorize Text.

Stemming: A group of different words may share the same word stem. A text retrieval system needs to identify groups of words where the words in group are small syntactic variants of one another and collect only the common word stem per group. For example the group of words drug, drugged and drugs, share a common word stem, drug and can be viewed as same occurrences of the same word. If, for instance, a searcher enters the term stemming as a part of a question, it's possible that he or she's going to even be inquisitive about such variants as stemmed and stem.. Conflation are often either manual-using some quite regular expressions or automatic, via programs referred to as stemmers. Stemming is additionally utilized in IR to cut back the dimensions of index files. Since one stem usually corresponds to many full terms, by storing stems rather than terms, compression factors of over 50% are often achieved. Version of the Porter stemming algorithm with a few changes towards the end in which we have omitted some cases e.g.:

- Prepare-Prepared-Prepares-Preparing
- Print-printing-prints-printed

In both the cases, all words of the first example will be treated as 'Prepare' and all words of the second example will be treated as 'print'.

Stemming refers to identifying the roots of a certain word. Figure 7 depicts the stemming process. There are basically two types of stemming techniques, one is inflectional and other is derivational.

Relevant documents
Relevant and retrived documents
Retrived documents

Fig. 7: Relationship between the set of relevant documents and the set of retrieved documents

Derivational stemming can create a new word from an existing word, sometimes by simply changing grammatical category (for example, changing a noun to a verb) (Wikipedia). The type of stemming we were able to implement is called Inflectional Stemming. A commonly used algorithms is the 'Porter's Algorithm' for stemming. When the normalization is confined to regularizing grammatical variants such as singular/plural or past/present, it is referred to inflectional stemming. To minimize the effects of inflection and morphological variations of Words (stemming), our Approach has pre-processed each word using a provided The advantage of stemming at assortment time is potency and index file compression-since index terms square measure already stemmed, this operation needs no resources at search time and therefore the index file are compressed as represented higher. The disadvantage of assortment time stemming is that info regarding the total terms are lost, or extra storage are needed to store each the stemmed and unstemmed forms.

Document visualization: There are a good number of text Information Retrieval products that fall into this category. The general approach is to organize the documents based on their similarities and present the groups or clusters of the documents in certain graphical representation. The subsequent register is by no means exhaustive but is enough to demonstrate the variety of the representation schemes available. Cartia's them escape is an enterprise information mapping application that presents clusters of documents in landscape demonstration. Canis'sc Map is a file clustering and visualization tool based on Self Organizing Map. IBM's Technology looks at, developed together with Synthema in Italy and is a text Information Retrieval request in the scientific field. It performs text clustering plus visualization in the shape of maps for patent databases and technical publications. Inxight also offers a visualization device, known as VizControls, which performs value added post processing of search results by clustering the documents into groups and displaying based on a hyperbolic tree representation.

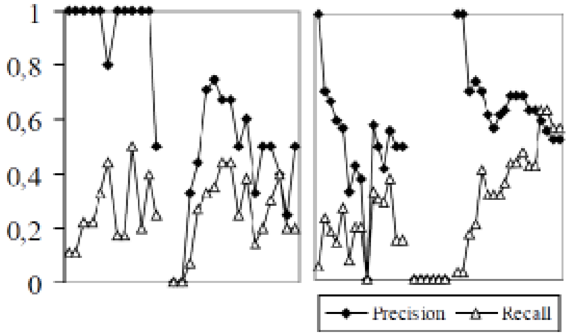

Fig. 8: Precision and recall for two

Semio Corp's SemioMap employs a 3D 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. Figure 8 depicts the UML activity in which the client provides the ability to define a customized IR-process as a UML-activity diagram. The server is responsible for managing the execution of defined IR-processes and for providing information about the available processor and datamodules. Every processor module provides a particular service functionality. These processor modules are separately invoked by the server for processing single steps in the context of the whole IR-process. The data modules hold input/output data for the processor modules and are also responsible for generating visualizations of their associated data content.

TEXT ANALYSIS AND UNDERSTANDING

Text analytics is still chiefly an undeveloped science and squeezes several different approaches. Text mining on the other hand is primarily concerned with the extraction of meaningful metrics from unstructured text data so they can be fed into data mining algorithms for pattern discovery.

Text mining platforms are a more recent phenomenon and provide a mechanism to discover patterns which might be used in operational activities. Text is used to generate extra features which might be added to structured data for more accurate pattern discovery. There is of course overlap and most suppliers provide a mixture of capabilities. Finally we should not forget information retrieval, more often branded as enterprise search technology, where the aim is simply to provide a means of discovering and accessing data that are relevant to a particular query. This is a separate topic to a large extent, although again there is overlap. The terms 'text mining', 'machine learning' and 'natural language processing' have different meaning depending on who you speak with. For our purposes 'text mining' is the application of algorithms to text data for the purpose of finding vulnerable patterns. 'Machine learning' is when a software system learns something so that a task can be performed more effectively next time around.

Alchemy API provides cloud based text analytics services to support sentiment analysis, marketing, content discovery, business intelligence and most tasks where natural language processing is needed. An on-site capability can also be provided if needed. The capabilities offered by Alchemy API go beyond those most large organizations could build in-house and not least because the training set used to model language is 250 times larger than Wikipedia. Innovative techniques using deep learning technologies (multi-layered neural networks) also go well beyond most of the competition and Alchemy API distinguishes itself by using the technology for image recognition in addition to text analytics. The functionality is broad and includes: Named entity recognition for identifying people, places, companies, products and other named items. Sentiment analysis with sophisticated capabilities such as negation recognition, modifiers, document level, entity, keyword and quotation level sentiment. Keyword extraction to identify content topics. Concept tagging, which is capable of identifying concepts not explicitly mentioned in a document. Relation extraction where sentences are parsed into subject, action and object. Text categorization to identify most likely categories. Other functionality such as author extraction, language detection and text extraction. Alchemy API was founded in 2005 and is based in Denver Colorado. Pricing plans for the cloud based services are based on transaction per day and start with a free Starter subscription. Knowledge READER from Angoss is part of a broad suite of analytics tools and specifically addresses text analytics in the context of customer oriented and marketing applications. It majors on visual representation including dashboards for sentiment and text analysis and also provides a somewhat unique map of the results of association mining to display words that tend to occur together.

Many of the advanced features make use of the embedded Lexalytics text analytics engine-widely recognized as one of the best in class. Entity, theme and topic extraction are supported along with decision and strategy trees for profiling, segmentation and predictive modeling. Sentiment analysis supports the visual graphing of sentiment trends and individual documents can be marked up for sentiment. Angoss provides its technology through the cloud or on-site implementation. High levels of end-user functionality are claimed with much of the functionality available to business users. More advanced analysis can be achieved by combining text with structured data and text can be used to generate additional features for data mining activities. Obviously this is a sophisticated product best suited to the needs of large organizations in the main, although the cloud based access will suit the needs of some mid-sized organizations too. Overall this is well suited to customer and marketing text analytics needs where text is used to gain insight into sentiment and customer behavior. Attensity majors on social analytics, but also offers a general purpose text analytics engine. Four major components define the offering; Attensity Pipeline collects data from over one hundred million social sources as input for analysis. Attensity Respond provides a mechanism for responding to social comment. Attensity Analyze allows text in emails, call-center notes, surveys and other sources of text to be analyzed for sentiment and trend. Attensity Text Analytics provides an underlying engine that embraces several unique NLP technologies and a semantic annotation server for auto-classification, entity extraction and exhaustive extraction. It comes with good integration tools too so that the results of text analytics can be merged with structured data analytics. Three horizontal solutions are offered for marketing, customer service and IT (Fig. 9).

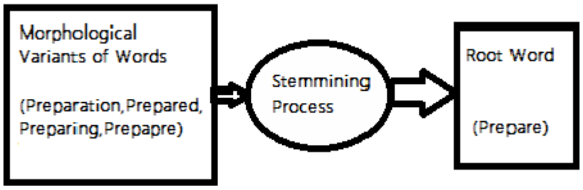
Morphological Variants of Words
(Preparation,Prepared, Preparing,Prepapre)
Stemmining Process
Root Word
(Prepare)

Fig. 9: The stemming process

Eaagle provides text mining technology to marketing and research professionals. Data is loaded into Eaagle and a variety of reports and charts are returned showing relevant topics and words, word clouds and other statistics. Both online and Windows based software is offered. The Windows offering is called Full Text Mapper with good visuals to explore topics and various word statistics. Expert System majors on semantic analysis, employing a semantic analysis engine and complete semantic network for a complete understanding of text, finding hidden relationships, trends and events and transforming unstructured information into structured data. Its Cogito semantic technology offers a complete set of features including: semantic search and natural language search, text analytics, development and management of taxonomies and ontology, automatic categorization, extraction of data and metadata and natural language processing. At the heart of Cogito is the Sensigrafo, a rich and comprehensive semantic network, which enables the disambiguation of terms, a major stumbling block in many text analytics technologies. Sensigrafo allows Cogito to understand the meaning of words and context (Jaguar: car or animal, apple: the fruit or the company?) -a critical differentiator between semantic technology and traditional keyword and statistics based approaches. Sensigrafo is available in different languages and contains more than 1 million concepts,

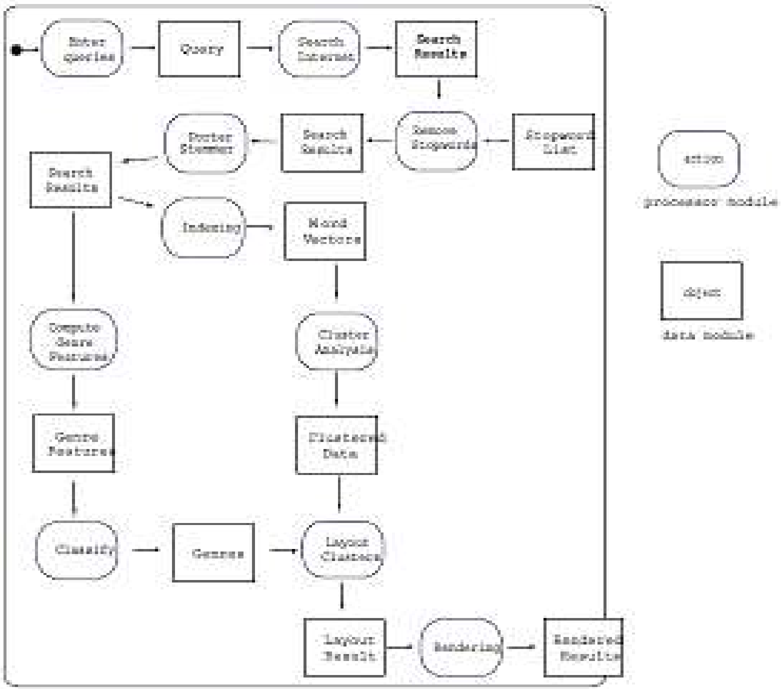

Fig. 10: The UML activity diagram shows all necessary actions and objects for the specific visualization process

more than 4 million relationships for the English language alone and a rich set of attributes for each concept. The Cogito semantic network includes common words, which comprise 90% of all content and rich vertical domain dictionaries including Corporate and Homeland Security, Finance, Media and Publishing, Oil and Gas, Life Sciences and Parma, Government and Telecommunications, providing rich contextual understanding that improves precision and recall in the process information retrieval and management. The technology has found uses in CRM applications, product development, competitive intelligence, marketing and many activities where knowledge sharing is critical. IBM provides text analytics support through two products. IBM Content Analytics is primarily an extension of enterprise search technologies that adds several useful visualizations to discover structure within text data. Langauge Ware on the other hand leverages Natural Language Processing (NLP) to facilitate several types of text analysis. Statistica Text Miner is part of the extensive Statistica statistical analysis and data mining product set. Extensive pre-processing options are available with stemming and stub lists for most European languages. 'Bag of words' type analysis can be carried out with input to the data mining capabilities of Statistica (Fig. 10).

DISCUSSION AND RESEARCH DIRECTION

There are still several critical research problems that need to be solved before frequent pattern mining can become a cornerstone approach in data mining applications. First, the most focused and extensively studied topic in frequent pattern mining is perhaps scalable mining methods. The set of frequent patterns derived by most of the current pattern mining methods is too huge for effective usage. However, it is still not clear what kind of patterns will give us satisfactory pattern sets in both compactness and representative quality for a particular application and whether we can mine such patterns directly and efficiently. Much research is still needed to substantially reduce the size of derived pattern sets and enhance the quality of retained patterns. Second, although we have efficient methods for mining precise and complete set of frequent patterns, approximate frequent patterns could be the best choice in many applications. Much researches is still needed to make such mining more effective than the currently available tools in bioinformatics. Third, to make frequent pattern mining an essential task in data mining, much research is needed to further develop pattern based mining methods. Fourth, we need mechanisms for deep understanding and interpretation of patterns, e.g., semantic annotation for frequent patterns and contextual analysis of frequent patterns. The main research work on pattern analysis has been focused on pattern composition (e.g., the set of items in item-set patterns) and frequency. In many cases, frequent patterns are mined from certain data sets which also contain structural information. Fifth, the most obvious recent example of this type of change is the rapid growth of mobile devices and social media. One response from the IR community has been the development of social search, which deals with search involving communities of users and informal information exchange. New research in a variety of areas such as user tagging, conversation retrieval, filtering and recommendation and collaborative search is starting to provide effective new tools for managing personal and social information. Sixth, 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. Seventh, most text mining tools focus on processing English documents, mining from documents in other languages allows access to previously untapped information and offers a new host of opportunities. Eighth, Domain knowledge could also play a part in knowledge distillation. It is also interesting to explore how a user's knowledge can be used to initialize a knowledge structure and make the discovered knowledge more interpretable. Ninth, Text mining tools could also appear in the form of intelligent personal assistants. 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. Information filtering is the process to help people find the valuable information.

CONCLUSION

Text mining research aims at improving human's ability to figure out huge data to provide better insights by analyzing a document or documents. It is time to re-examine the evaluative methods that are used information retrieval studies, particularly with respect to search performance on the Web. New measures can be defined that take into account hyper linking and relevance ranking. Discovering such hidden knowledge is an essential requirement for many corporations, due to its wide spectrum of applications. In this short survey, the notion of text mining has 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, a non-HTML 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. More extensive studies are needed to assess the properties of different search engines and evaluative measures, particularly in realistic search tasks.

REFERENCES

Ai, Y., R. Gerling, M. Neumann, C. Nitschke and P. Riehmann, 2005. TIRA: Text based information retrieval architecture. Proceeding of the 2nd International Workshop on Text-based Information Retrieval, pp: 345-352.

Barathi, M. and S. Valli, 2011. Context disambiguation based semantic web search for effective information retrieval. J. Comput. Sci., 7(4): 548-553.

Preethi, M., 2014. Concept based ontology matching by concept enrichment. Proceeding of International Conference on Global Innovations in Computing Technology (ICGICT'14), pp: 234-243.

Sagayam, R., S. Srinivasan and S. Roshni, 2012. A survey of text mining: Retrieval, extraction and indexing techniques. Int. J. Comput. Eng. Res., 2(5): 1443-1444.

Vasumathi, B. and S. Moorthi, 2012. Implementation of hybrid ANN-PSO algorithm on FPGA for harmonic estimation. Eng. Appl. Artif. Intel., 25: 476-483.

Wang, Y.F., S. Cheng and M.H. Hsu, 2011. Incorporating the Markova chain concept into fuzzy stochastic prediction of stock indexes. Appl. Soft Comput., 10(2): 613-617.