Web Service to Execute A Datamining Task

R.Velkumar  
Research Scholar, Dept. of Computer Science, VELS University, Chennai

Dr.A.Muthukumaravel  
Professor & Head, Department of MCA, BHARATH University, Chennai.

N.Sathya  
Assistant Professor, Dept. of MCA, VELS University, Chennai.

Abstract - Data mining is the process of extracting hidden patterns from large amounts of data. Data mining is becoming an increasingly important tool to transform this data into information. This system proposes for multi document summarization by using an approach to address the problem of improving content selection (Extraction), in automatic text summarization by using some methods in fuzzy concepts. In proposed system the corpus (Dataset) is collected from the online search engine. The system explores the problem of subtopic segmentation by proposing a fuzzy model for the semantics of both words and sentences. Single document summarization is done by using sentence scoring. The Single Document Summary outputs of similar documents are used as an input for Multi Document Summarization. Sentence Similarity is used to compare the sentences and the resultant summary is produced with Naïve Bayes Classifier.

Keywords: Web service, Data mining, Fuzzy logic and Naïve baves.

I. INTRODUCTION

Data mining is the process of extracting hidden patterns from large amounts of data. Data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.

Data mining can be applied to data sets of any size. However, while it can be used to uncover hidden patterns in data that have been collected, it can neither uncover patterns that are not already present in the data, nor can it uncover patterns in data that have not been collected.

1.1 DATA MINING PROCESS

Various kind of data mining process is available. The Data Mining process is shown in the Figure 1.

1.2 TEXT MINING

Text mining sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text [7]. High-quality information is typically derived through the divining of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment
II. LITERATURE SURVEY

Investigated the correlation between ROUGE and human evaluation of extractive meeting summaries and focus on a few issues specific to the meeting domain in [1]. Both human and system generated summaries are used. We conducted human evaluation of different summaries, calculated ROUGE scores with several variations, and examined their correlation with human evaluation based on Spearman’s rho. Addressed the design and implementation of Bilingual Information Retrieval system on the domain, Festivals. A generic platform is built for Bilingual Information retrieval which can be extended to any foreign or Indian language working with the same efficiency. This paper deals with Indian language Tamil apart from English [3]. Proposed Web-based metrics that compute the semantic similarity between words or terms are presented and compared with the state of the art. The proposed algorithms work automatically, do not require any human-annotated knowledge resources, e.g., ontologies, and can be generalized and applied to different languages [4].

Presents the keyword extraction techniques, exploring the effects that part of speech tagging has on the summarization procedure of an existing system. Moreover, author describes the profiling features that are used as an extension to an already constructed news indexing system, PeRSSonal. Author thus enhancing the personalization algorithm that the system utilizes with various features derived from the user’s profile, such as the list of viewed articles and the time spent on them [5]. Utilized a Hidden Markov Model (HMM)-based bilingual (Mandarin and English) text-to-speech (TTS) system to synthesize natural speech for given bilingual text. A new, mixed language TTS is constructed in decision-tree based clustering. By sharing states, the new system has a smaller footprint than the baseline system. The nearest neighbor is measured with the Kullback–Leibler divergence (KLD) and mappings between leaf nodes in the decision trees of the source and target languages are established. An average voice based on HMMs can be first trained by using speech data of speakers in different languages [6].

2.1 EXISTING SYSTEM

- In Single Document summarization the summary of only one document can be built.
- Shallow text processing approaches as opposed to semantic approaches related to natural language processing.
- automatic generation of text summaries which includes:
  - Extraction of multiword descriptions of single document.
  - Extraction of multiword descriptions of collection of documents.

2.2 PROPOSED SYSTEM

- In Multi Document summarization the summary of multiple documents can be built.
- ROUGE based method is proposed to summarize the multi documents.
- The summarization methods can be classified into abstractive and extractive summarization.
- Text abstraction and text extraction. Text abstraction examines a given text using linguistic methods which interpret a text and find new concepts to describe it. And then new text is generated which will be shorter with the same content of information. Text extraction means extract parts (words, sequences, sentences, paragraphs, etc.) of a given text based on statistic, linguistic or heuristic methods, and then join them to new text which will be shorter with the same content of information.
- A typical extractive summarization method consists in several steps, at each of them different options can be chosen. We will assume that the units of selection are sentences (these could be, say, phrases or paragraphs). Thus final goal of the extractive summarization process is sentence selection.
- A summary of a document is a (much) shorter text that conveys the most important information from the source document. There are a number of scenarios where automatic construction of such summaries is useful. For example, an information retrieval system could present an automatically built summary in its list of retrieval results, for the user to quickly decide which documents are interesting and worth opening for a closer look.

III. FUZZY LOGIC

Fuzzy logic is a powerful problem-solving methodology with a myriad of applications in embedded control and information processing. Fuzzy provides a remarkably simple way to draw definite conclusions from
vague, ambiguous or imprecise information. In a sense, fuzzy logic resembles human decision making with its ability to work from approximate data and find precise solutions.

Unlike classical logic which requires a deep understanding of a system, exact equations, and precise numeric values, Fuzzy logic incorporates an alternative way of thinking, which allows modeling complex systems using a higher level of abstraction originating from our knowledge and experience. Fuzzy Logic allows expressing this knowledge with subjective concepts such as very hot, bright red, and a long time which are mapped into exact numeric ranges.

Fuzzy Logic is basically a multivalued logic that allows intermediate values to be defined between conventional evaluations like yes/no, true/false, black/white, etc. Notions like rather warm or pretty cold can be formulated mathematically and processed by computers.

**MEMBERSHIP FUNCTION**

The membership function is a graphical representation of the magnitude of participation of each input. It associates a weighting with each of the inputs that are processed, define functional overlap between inputs, and ultimately determines an output response. The rules use the input membership values as weighting factors to determine their influence on the fuzzy output sets of the final output conclusion. A graph that defines how each point in the input space is mapped to membership value between 0 and 1. Input space is often referred as the universe of discourse or universal set (u), which contain all the possible elements of concern in each particular application.

**Features of Membership Function**

![Membership Functions](image)

**IV. RESULTS**

The dataset (corpus) is collected on cricket newses from the search engines for test matches, one day matches and T20. English Documents are collected for the following series: India – South Africa, Australia – England, Pakistan – New Zealand from ESPN cricinfo. We selected Australia – England (Ashes) series as a training set. We use 200 documents for training set and 50 documents for test set. The test series includes 5 test matches. Each match is conducted for 5 days. The reports are gathered from day1 to day5 for all five test matches. Every match includes five documents or reports. Reports are segmented into subtopics by using a fuzzy model. Document segmentation is based on some important events occurred in the match. (Example: Run Scoring Fours / Sixes, Wickets, Debutants, Records).

Segmented single document topics are summarized by using sentence scoring. The resultant summary will be provided the consolidated report for each test match. The output summary of five test matches is used as an input of the MDS. Sentence Similarity is used to compare the sentences and the resultant summary is produced with Naive Bayes Classifier. The resultant summary will be provided the consolidated report for test series.
One day match and T20, the information is collected from all matches of the series. Single match report is used as an input for Single Document Summarization. All Single Document Summaries of one day matches or T20 matches are used as an input for Multi Document Summaries. The resultant summary is provided the consolidated report for the particular series.

The results are plotted on graph as shown in Figure 3.

Fig.3 Comparison of Baseline system with Proposed System for

REFERENCES


