NAME:Aditya Somani CLASS:TE-A ROLL NO:T1851061

Case Study : Text Mining

Introduction

It is the uncoveringand unveiling the hidden patterns by the use of computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation.

Text mining (TM) seeks to extract useful information from a collection of documents. It is similar to data mining (DM), but the data sources are unstructured or semi-structured documents.The TM methods involve :

- Basic pre-processing / TM operations, such as identification / extraction of representative features(this can be done in several phases)
- Advanced text mining operations, involving identification of complex patterns(e.g. relationships between previously identified concepts)TM exploits techniques / methodologies from data mining, machine learning, information retrieval,corpus-based computational linguistics.

Need

While traditional search engines like Google now offer refinements such as synonyms, auto-completion and *semantic search* (history and context), the vast majority of search results only point to the location of documents, leaving searchers with the problem of having to spend hours manually extracting the necessary data by reading through individual documents.

The limitations of traditional search are compounded by the growth in big data over the past decade, which has helped increase the number of results returned for a single query by a search engine like Google from tens of thousands to hundreds of millions.

To solve these issues text mining can be used. If we categorize the gathered big data based on keywords we can easily navigate through it and utilize it properly and more efficiently.

Methods of Text Mining

Data Mining Style: View text as high dimensional data

- Frequent pattern finding
- Association analysis
- Outlier detection

Information Retrieval Style: Fine granularity topical analysis

- Topic extraction
- Exploit term weighting and text similarity measures
- Question answering

Natural Language Processing Style: Information Extraction

- Entity extraction
- Relation extraction
- Sentiment analysis

Machine Learning Style: Unsupervised or semi-supervised learning

- Generative models

- Dimension reduction
- Classification & prediction

Applications

- Spam filtering
- Monitoring public opinions (for example in blogs or review sites)
- Customer service, email support
- Automatic labeling of documents in business libraries
- Fraud detection by investigating notification of claims
- Analyzing open-ended survey responses
- Automatic processing of messages, emails, etc.
- Investigating competitors by crawling their web sites.

Advantages

Marketing / Retail

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc.

Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank, and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner

Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipment and determine optimal control parameters.

Governments

Data mining helps government agencies by digging and analyzing records of the financial transaction to build patterns that can detect money laundering or criminal activities.

Disadvantages

Privacy Issues

The concerns about personal privacy have been increasing enormously recently especially when the internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are afraid that their personal information is collected and used in an unethical way that potentially causes them a lot of trouble.

Security issues

Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc.

Misuse of information/inaccurate information

Information collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take advantage of vulnerable people or discriminate against a group of people.