Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 09, September-2019

Survey on Ranking Concepts and Text Mining Algorithms

Ms. N. Banupriya
Assistant Professor
Department of Computer Science and Engineering
R.M.K. Engineering College
Chennai, India

Dr. T. Sethukarasi
Professor and Head
Department of Computer Science and Engineering
R.M.K. Engineering College
Chennai, India

Ms. A. Jasmine Gilda
Assistant Professor
Department of Computer Science and Engineering
R.M.K. Engineering College
Chennai, India

***Abstract*— The Problem arises in day to day surfing is all about classification of text (i.e.) Separation of documents and serving according to hierarchy level. The person one who uses internet should be provided with information what is requested by him/her. This separation of information is based on the request, usage of past users, repeated usage of individuals etc. There occurs some problem in news filtering and organizing of text, document organization and retrieval, email classification, spam filtering which is subjected mainly on text classification or data classification. These all can be solved using Ranking algorithms which categorizes the documents based on text mining concepts or key for example by giving score or mark for each document which is surfed daily. This paper deals with scoring the documents efficiently by Ranking algorithms and relate how the ranking concepts come in real world.**

Keywords—Text Classification, Ranking, Documents, Filtering

I. INTRODUCTION

Data deals with mining of data from warehouse where the information about data is stored. Text mining is used to find the hidden patterns from data mining Text mining[1] covers large number of areas and mainly it deals with Information Extraction, Natural Language Processing, Concept Extraction, Web mining, Document clustering, Information Retrieval. There are many several Text mining algorithms for ranking .The current research areas are Named Entity Recognition, Relationship Extractions, Text Classification, Synonym and abbreviation Extraction.

The Fig 1. shows the components of Text classification 1] or Text analytics in which Ranking Play a main role to categorize a text from data and rank it in order give the effective reply for the request made by user.

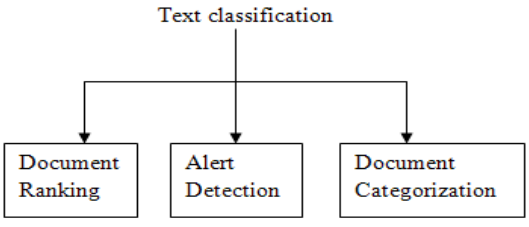

Fig1. Text Classification Components

II. INFORMATION RETRIEVAL

Information Retrieval [2], [11], also plays a major role in ranking where it is also the cause of ranking concepts. In Text mining, IR Deals with these areas such as Keyword Search/Querying, Indexing, Document Ranking, Page Rank, and Document Similarity. Page ranking Concept is common to Web mining and IR, Document Ranking is common to Web mining, IR and Document Classification, Document Similarity is common to Document Clustering and IR

A. First Machine for Text Surfing:

In 1948, A Convened Conference held by UK Royal Society, in that Holmstron Described "Machine called the UNIVAC "[3] is capable of searching for text references associated with subject code, where text and subject code are associated with magnetic steel tape. In this Machine process 120 words per min is trained and it is the first reference to a computer to search content (Text or Data).Mitchell described UNIVAC that it will search 1000000 record indexed by up to six subject codes. It would take 15h to search many records. Nanus says computed based IR projects run in 1950 includes one system for general electric that searched over 3000 document abstracts. Based on this another work was conducted in the Soviet Union in 1950.

These researches reach other field and Hollywood drew public attention on IR with comedy desk set in 1957. This helps to refer libraries to be replaced by a computer. Later IR Became a research areas which deals with how to index, how to store and how to retrieve items.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 09, September-2019

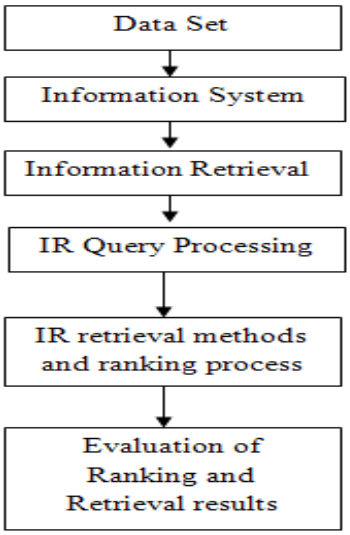

Fig 2. Information Retrieval Ranking System

III. RANKING

Boolean Retrieval [3],[7] is the combination of Electromechanical and computer based IR Systems. Luhn and Maron says that each document is assigned with score ,it indicates its relevance to that query .The documents are sorted and the scores are applied to the matches with the existence. Some manual keywords are assigned to a collection of 200 or 190 documents and weight is assigned according to the matches observed. Effective ranking of documents or text will give formatted indexing to refer in real world.

Ranked Retrieval method says 39 Queries are used and it wins the Native method of Boolean Search. Later they furnish a useful measurement of word significance about occurrence of word. This Ranked Retrieval application was later taken by IR Researches to search, redefining usage to query relationship over those years. The Fig 2 Shows the Ranking system of IR.

A. Approaches in Ranking

- Content based Ranking
- Usage based Ranking
- Link based Ranking
  - (i) Page ranking Algorithms
  - (ii) Hits Algorithm

Units Overall Query is computed by summarizing of all the weights of key word and content in content based ranking technique [4]. Query is given and it is analyzed based on keywords and content obtained. Now preprocessing is done to identify the root words, it constructs the dictionary which will provide the meaning for root words. After capturing content keywords and page keywords weight is assigned to each word.

Based on user usage of pages in internet the ranking is provided and according to the current visit and past users it is decided. Structural Properties of the web graph are ignored. Mainly based on saving, printing, adding the page, actions taken in the page is all good indicators for the usage based ranking [4], [6].

Calculation of this ranking algorithm is independent of Query. Linking the structure of document is called Linked based ranking [4], [5] . The Ranking is based on Citations, References of text in several pages must be calculated static before getting Query. To maintain in structure format ranking algorithm are applied for web page. In this content score, web mining techniques are used to order them.

IV. LITERATURE SURVEY

[2] In this paper, they described about IR usage for ranking the documents and some ranking algorithms are Indegree, Hits, Page rank, SALSA .These algorithms are based on Query processing to attain the accurate result in classification.

In [4], [10] the steps to perform score for a page is described .Every node are assigned between o and 1 at first. This is the basic technique in Page rank and depends on how much link structure is made in the web graph. The score for each text in the web page are calculated using cosine similarity; Probability features of text, term proximity were all involved. This final score is used to list the documents and will be results of query. Page ranking involves Markov process and computation of steady state distribution for calculating the rank of text or documents

HITS [4], [9] are associated with Hubs and Authorities. It does not include Query language. Two scores are given in HITS, one hub score and Authority score from which finally rank is produced. In Usage based ranking [6], Semantic relationship analyzes the contents of web pages. In that Ontology is used to analyze the content relationship. Link relevant is prioritized based on term weight. Term weight is assigned based on their search history weight of the user.

Rocchio is a Text classification algorithm where it is a relevance feedback method.

Indexing [3] is debated in the field of Librarianship and it is a classic approach. All the contents are placed in hierarchy level using Decimal system in order to represent using numerical codes. List of keywords in uniterm system is proposed by Taube et al. Detailed Retrieval effectiveness and some classic classification Techniques for uniterm system was described by Cleverdon. He used Words to index the document of IR system and his test collection approach is still used for academic research and Commercial testing purpose today. Later on the ranking concepts is started with Boolean Retrieval concepts.

In [8] Page rank algorithm is explained along with Page rank modification, HITS, Hilltop, Query sensitive page ranking etc. These ranking algorithms concentrate on ranking the search engine by classification of text, combining user's feedback, web page ranking on trust based and similarity.

[9] In this paper they concentrated on top 10 frequently used algorithm, their brief justification and representative publication reference. They analyze with the results generated .C4.5 and beyond algorithm construct classifiers tool in Data mining. The input given here is collection of cases (small number of classes) and output is a classifier that can accurately predict the class to which a new case belongs. K-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 09, September-2019

means algorithm is a simple iterative method.SVM insists on finding the maximum margin hyper planes are that it offers the best generalization ability for classification performance on training data. Page rank algorithm is a search ranking algorithm using hyperlinks on the web in text classification method and some more algorithms are also analyzed in this paper.

V. TEXT MINING ALGORITHMS USED CURRENTLY

Algorithms used for Text classification [12] are Bag of words (manual approach), Statistical systems and rule based systems. Algorithms used for Named Entity Recognition are Rule based approach (set of Rules) or through Machine learning and Statistical learning approach (translates into a sequence labeling problem).Algorithms used for Relationship Extractions are featured based classification and Kernel method (used in machine learning).

Generally Text mining algorithms[12] are K-nearest Neighbor algorithm, Rocchio Algorithm, Decision Tree, Rule Learning and some machine learning algorithms are also concentrated along with Text classification algorithms like Naive Bayes classifier, Support SVM, Logit Boost, LSI, Radial basis Function Networks. These are some algorithms which are used for specific action in Part of speech tagging, Link analytics, Keyword search, Inverted index, Word representation etc.

A. Analysis of Algorithms

Algorithms	Deals with	Advantages
NAIVE BAYES	Statistical Algorithms based on Bayes's theorem which helps to compute probabilities	Used in Machine Learning
SUPPORT VECTOR MACHINES	Vector Representation and achieve more accurate results with computational recourses.	Used in Machine Learning
DEEP LEARNING	Set of data's trained according to the human brain work which deals with Convolutional Neural Networks (CNN) architecture and Recurrent Neural Networks (RNN) architecture. Results are more accurate.	Used in Machine Learning
K-NEAREST NEIGHBOR (KNN)	Compare the existing text with classified new texts. Its application is Social media monitoring etc. Simple and easy	Used in sentiment analysis
K-MEANS CLUSTERING	Classical algorithm which transform text to numerical data.	Used in Document ranking and document classification
DECISION TREE	It has tree structure with root, leaf and branch nodes. It analyses the result in each branch node, attributes are internal node and class label is leaf node.	Used in Machine Learning
GENERALIZED LINEAR MODELS (GLM)	Statistical algorithms based on linear model includes regression model.	Used in Machine Learning
APRIORI ALGORITHM	Classical Algorithm which specifies association rules to determine relationships with large database	Used in E-Commerce applications
PAGERANK	It relates the Quality and Quantity of Web links used, it has set of rules to be framed in filtering the spam etc	Used in text mining and web mining
NON-NEGATIVE MATRIX FACTORIZATION	It is Linear Algebra which deals with matrices.	Used in text mining ,Spectral data analysis
MINIMUM DESCRIPTOR LENGTH	Statistical calculations and Probability theory. Bayes theorem also considered	Can apply in Machine Learning
BOOSTING (LP boost, logit boost, xgboost)	It weights the training data with respect to Phenomenon.	Used in Machine Learning and Computational Learning theory

Fig 3 Analysis of Algorithms in Text mining

VI. CONCLUSION:

Text classification deals with number of problems like word relations, dependency relations, context relations, word representations etc. these all concepts can be reviewed with 'n' number of algorithms in text mining. In this paper we concentrate on ranking concepts, latest used text mining algorithm in current areas. Among Ranking algorithm, page ranking algorithm[10] is an algorithm which is the origin of Google ranking. Many researchers have modified this and several algorithms is developed according to their concepts and it is still under progress for efficient results to be obtained.

REFERENCES

[1] "The Seven Practice Areas of Text Analytics Excerpt from: Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications" G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, and R. Nisbet, Elsevier, January 2012 Available now: http://amzn.to/textmine.

[2] "A Review: Ranking documents using Ranking Algorithms & Techniques", Rana Azhar-ul-Haq (12031719-020) University Of Gujrat, Punjab Pakistan azhar4u@live.com Date· May 2013, https://www.researchgate.net/publication/

[3] "The History of Information Retrieval Research", in Proceedings of the IEEE 100(Special Centennial Issue):1444-1451 · May 2012 Mark Sanderson: School of Computer Science and Information Technology, RMIT University, W. Bruce Croft: Department of Computer Science.

[4] "Survey on Different Ranking Algorithms Along With Their Approaches", International Journal of Computer Applications (0975 – 8887) Volume 135 – No.10, February 2016, Nirali Arora, Sharvari Govilkar Department of Computer Engineering. PIIT Mumbai University, India.

[5] Azam Feyznia,mohsin Kahanti "A link analysis based ranking method for semantic web documents "at ieee proceedings of 2010.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 09, September-2019

[6] "Survey on Web Page Ranking Algorithms", International Journal of Computer Applications (0975 – 8887)Volume 41– No.19, March 2012,Mercy Paul Selvan M.E, Department of Computer Scienc Sathyabama University, A .Chandra Sekar M.E Ph.D,Department Of Computer Science St.Joseph's College of Engineering, A.Priya Dharshin Department of Electronic Science Sathyabama University.

[7] "Web search basics", April 1, 2009 Cambridge University Press, Online edition 2009 Cambridge UP.

[8] "An Overview of Ranking Algorithms for Search Engines",in Proceedings of 2nd National Conference,INDIACom-2008,Feb 8,9,byAnkur Gupta,PG Scholar,Delhi College of Engineering RajniJindal,Faculty, Delhi College of Engineering,Delhi.

[9] Top 10 algorithms in data mining, Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg, 8 October 2007 .

[10] "The PageRank Citation Ranking: Bringing Order to the Web ",January 29, 1998

[11] "An Introduction to Information Retrieval", Christopher D. Manning Prabhakar Raghavan Hinrich Schütze, Cambridge University Press Cambridge, England, Online edition (c) 2009 Cambridge UP. Printed on April 1, 2009 Website: http://www.informationretrieval.org/ Comments, corrections, and other feedback most welcome at: *informationretrieval@yahoogroups.com*.

[12] Text mining Algorithms in Expert System Semantic Intelligence ,Feb 7,2017,http://www.expertsystem.com/text-mining-algorithms/

[13] http://intellspot.com/text-mining-algorithms/.