International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.03, April 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

SoWeRank:

Hybrid Approach For Searching and Ranking Large Scale Web Data Using Social Media Factors

Yogesh Lonkar[1], Dattatray Savant[2], Prashant Nikam[3], Suraj Halkude[4], Kamesh Patil[5]

[1,2,3,4,5] *Genba Sopanrao Moze College of Engineering, Balewadi 411045(India)*

ABSTRACT

In last few years, there has been a hard development of incorporating results starting structured data source into keyword based web track systems such as Amazon as well as Google or any search engines. In search engines, different users may seek for different information by issuing the similar query. To convince more users with partial search results, search result diversification re-ranks the results to coat as many user intents as probable. Most presented intent-aware diversification algorithms differentiate user intentions as subtopics, every of which is typically a word, a phrase, or a piece of clarification. Web search queries are often uncertain or multi-search, which makes a effortless ranked list of results insufficient. To help information finding for such queries, system explore a technique that explicitly represents fascinating meaning of a query using groups of semantically related terms retrieved from search results. In the proposed work system denote a supervised approach based on a graphical model to identification of web queries show that the supervised approach significantly outperforms existing methods, which are mostly unsupervised, recommending that query facet retrieval can be effectively learned. First, the temporal prevalence of a particular topic in the news media is a factor of importance, and can be considered the media focus of a topic. Second, the temporal prevalence of the topic in social media indicates its user attention. Last, the interaction between the social media users who mention this topic indicates the strength of the community discussing it, and can be regarded as the user interaction toward the topic. We propose an unsupervised framework approach which identifies any type of search query, and then ranks them by relevance using their weight as well as their number of visit by users.

Keywords: Information Filtering, Social Computing, Social Network Analysis, Topic Identification, Topic Ranking

I. INTRODUCTION

1.1 Overview

With the currently growing interest in the Semantic Web, it is reasonable to expect that increasingly more metadata describing domain information about resources on the Web will become available. The idea presented here is to enrich the search process for hypermedia applications with information extracted from the semantic model of the application domain. One of the novelties in the semantic search proposed is the combination of spread activation techniques with traditional search engines techniques to obtain its results. One of the greatest

problems of traditional search engines is that they typically are based in keyword processing. Consider the following motivating example for a research institution domain. This domain deals with people, publications and research areas. Notice that "Keyword" is not a concept of the model, but is used in the diagram to repress fact that a keyword occurs inside the textual representation of the associated concept instances. For instance, the keyword "web" occurs inside the concept instance "The Evolution of Web Services" since it appears in the publication's "title" property. The keyword "ontology" is also related to the same concept since it appears in its "abstract" property.

A query with the keyword "web" would have as results only nodes of type Publication where this word occurs. If the user searches for nodes of type "Professor", the result could well be an empty set, since the keyword "web" may not appear inside the description text (page) of any of the professors. On the other hand, analyzing the semantics of this domain, it seems intuitive to think that if a given professor has many publications that are related to a given keyword, there is a great possibility that the professor himself is indeed related to that same keyword, and should be returned as a result of the query. Therefore, in the previous example, the Professor node "Schwabe" could be returned as a result for the query "web.

1.2 Background

Spread Activation techniques are one of the most used processing frameworks for semantic networks. It has been successfully used in several fields, particularly in Information Retrieval applications. Since it was developed in the Artificial Intelligence area as a processing framework for semantic networks and ontologies, we envision it to be a natural and interesting choice of knowledge processing algorithm in the context of the Semantic Web. An interesting and recent system that uses spread activation to process ontologies is ONTOCOPI, which tries to identify communities of practice within an organization. The spread activation algorithm works basically as a concept explorer. Given an initial set of activated concepts and some restrictions, activation flows through the network reaching other concepts which are closely related to the initial concepts. It is very powerful to perform proximity searches, where given an initial set of concepts, the algorithm returns other concepts which are strongly connected to them. An overview of spread activation techniques. Usually spread activation techniques are used either on semantic networks (where each edge in the network has only a label associated to it) or on associative networks (where each edge has only a numeric weight associated to it). In the Semantic Web, we use ontologies as the semantic network used by the spread activation algorithms.

II. LITERATURE SURVEY

In this section we cite the relevant past literature that utilizes the various techniques for ranking. Ranking search results is a fundamental problem in information retrieval. Most common approaches focus on the similarity of query and page as well as overall page quality. However, with increasing popularity of search engines the capturing of user behaviors insists to appear on surface. A lot of methods have been done on implicit measures of user preference in field of information retrieval.

We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. [1].

Probabilistic latent semantic analysis is a novel statistical technique for the analysis of two-mode and co-occurrence data, which has an application in information retrieval and filtering, natural language processing , machine learning from text, and in related areas[2].

Probabilistic Latent Semantic Indexing is a novel approach to automated document indexing which is based on a statistical latent class model for factor analysis of count data. [3].

We consider topic detection without any prior knowledge of category structure or possible categories. Keywords are extracted and clustered based on different similarity measures using the induced k-bisecting clustering algorithm. [4]

The clustering algorithm implemented in this system, called Induced Bisecting k-Means, outperforms the Standard Bisecting k-Means and is particularly suitable for on line applications when computational efficiency is a crucial aspect[5].

Then we describe the problems that are addressed in Statistical Natural Language Processing (NLP), like tagging and disambiguation, and a selection of important work so that students are grounded in the advances that have been made and, having understood the special problems that language poses, can move the field forward[6].

In this paper we recognize this primary role of Twitter and we propose a novel topic detection technique that permits to retrieve in real-time the most emergent topics expressed by the community. [7].

We use a Twitter-LDA model to discover topics from a representative sample of the entire Twitter. We then use text mining techniques to compare these Twitter topics with topics from New York Times, taking into consideration topic categories and types. We also study the relation between the proportions of opinionated tweets and retweets and topic categories and types. Our comparisons show interesting and useful findings for downstream IR or DM applications. [8]

To find topics that have busty patterns on microblogs, we propose a topic model that simultaneously captures two observations: (1) posts published around the same time are more likely to have the same topic, and (2) posts published by the same user are more likely to have the same topic. The former helps find eventdriven posts while the latter helps identify and filter out “personal” posts. Our experiments on a large Twitter dataset show that there are more meaningful and unique busty topics in the top-ranked results returned by our model than an LDA baseline and two degenerate variations of our model. [9].

To achieve prioritization, information must be ranked in order of estimated importance considering three factors. First, the temporal prevalence of a particular topic in the news media is a factor of importance, and can be considered the media focus (MF) of a topic. Second, the temporal prevalence of the topic in social media indicates its user attention (UA). Last, the interaction between the social media users who mention this topic indicates the strength of the community discussing it, and can be regarded as the user interaction (UI) toward the topic. [10]

III. RELATED WORK

News search is a technique for accessing information organized according to a faceted classification system, allowing users to digest, analyze and navigate through multidimensional data. It is widely used in e-commerce and digital libraries. Faceted search is similar to query facet extraction in that both of them use sets of coordinate terms to represent different facets of a query. However, most existing works for faceted search are build on as specific domain or predefined categories, while query facet extraction does not restrict queries in a specific domain, like products, people, etc.

For most people, the way they interact with web search engines has not changed significantly in the last decade. They still issue queries manually and review lists of result documents. The most significant and obvious user interface changes were the introduction of verticals (e.g. images, videos, and news), query auto complete, and question answering (e.g. Google Knowledge Graph). However, most internet users are also acquainted with faceted search: any ecommerce website, any library and most catalogues of any kind employ this technique to provide an accessible and fast way to locate arbitrary objects. System believe that most users would appreciate the utilization of this idea in web search. However, this is no trivial task. The ultimate goal of Faceted Web Search is to support the user to accomplish his search task. Previous work focused on the idea of using existing taxonomies or on generating facets for an entire corpus offline after indexation. These approaches lack the adaptation to the document result space or the user intent, and are too narrow. System propose web search facets that automatically recognize different subtopics, partition the search result space evenly and exhaustively per subtopic, and still contain only a small number of terms.

However, these original facets cannot be directly adopted as subtopics. Since query facets are designed for splitting different facets of a query, they are usually far more fine-grained than traditional subtopics in diversification. There are some objectives to achieve the successful implementation of system.

1. Need to implement novel base recommendation approach suing user feedback or search sessions.
2. To extract the data from the search engine databases related to query, searched by the user and represent the search results in restructured manner.
3. To provide search results according to search goals of particular user.
4. System have to provide a service recommendation base on similarity score which is calculating using text similarity algorithm.
5. Improve the system accuracy using the clustering algorithm base on potential users.

Using hierarchical association algorithm finds the user interest and collaborate the filter clusters.

IV. PROPOSED SYSTEM DESIGN

4.1 Objectives

- Extract the data from Twitter, Google as well as YouTube from third party search engines.
- To deploy the proposed system with Public cloud
- To enhance the system security using security as well as privacy algorithms e.g. SQL injection and prevention, pattern matching apaches.

- Enhance the system final ranking using weight and maximum visited page is like hybrid system
- Session storage of search queries
- Third party API used for multiple pages extraction from search engines.

4.2 Problem Statement

In the proposed research work to design and implement a system than work as classify and re-rank all type of query events along with the current events as well as news. The Google API will provide the third party interface for communicate with search engine the classify the all data using machine learning approach and re-rank with page rank as well as click through algorithms. We will collect the data from Google API, YouTube as well twitter and rank all the news base on current user query.

4.3 Project Overview

The proposed system consolidate and rank the most prevalent topics discussed in both news media and social media during a specific period of time. The system framework can be visualized in Fig. 1. To achieve its goal, the system must undergo four main stages. 1) Preprocessing: Key terms are extracted and filtered from news and social data corresponding to a particular period of time. 2) Key Term Graph Construction: A graph is constructed from the previously extracted key term set, whose vertices represent the key terms and edges represent the co-occurrence similarity between them. The graph, after processing and pruning, contains slightly joint clusters

of topics popular in both news media and social media. 3) Graph Clustering: The graph is clustered in order to obtain well-defined and disjoint TCs. 4) Content Selection and Ranking: The TCs from the graph are selected and ranked using the three relevance factors (MF, UA, and UI). Initially, news and tweets data are crawled from the Internet and stored in a database. News articles are obtained from specific news websites via their RSS feeds and tweets are crawled from the Twitter public timeline [41]. A user then requests an output of the top k ranked news topics for a specified period of time between date d1 (start) and date d2 (end).

4.4 Development Methodology

In this work first we extract the data from Twitter, Google as well as YouTube from third party search engines. Then extract the metadata from each page like title, meta tag as well description, so each page represent base on this extracted features. System used some training dataset for category classification or clustering. We apply any clustering algorithm for generate a similar cluster of news or web pages. Calculate each clusters average weight. Finally for ranking used cluster weight as well as visit count of each page. System will show the top k results base on twitted news, current Google news and videos also. After completion of system we presents some graphs which can define the system accuracy as well as time complexity.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.03, April 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

4.5 System Architecture

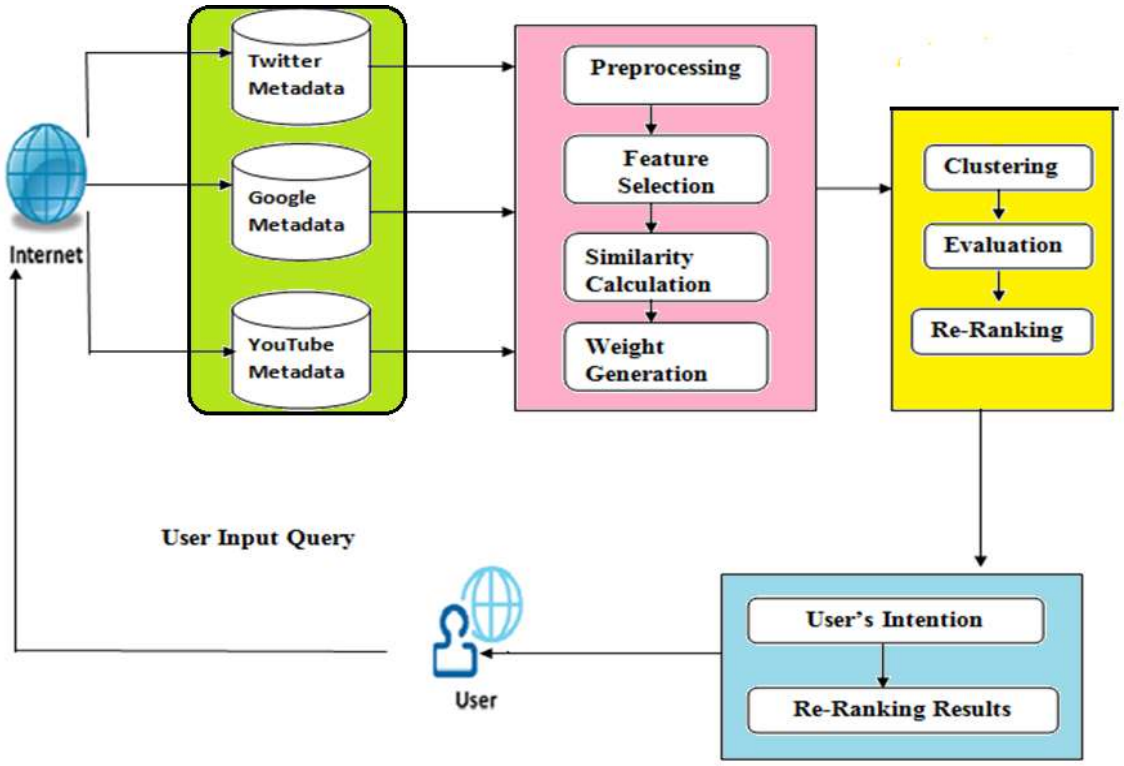

Figure 1 : Proposed System Architecture

4.6 Algorithms

4.6.1 Document retrieval Algorithm

Input: Users query as Q , Network Connection N;

Output: result from relevancy calculation top k pages base on Q.

Step 1: User provide the Q to system.

Step 2: if (N!=Null)

Process

Read each attribute A from ith Row in D

Res[i]=Calcsim(Q,A)

Else No connection

Step 3: For each(k to Res)

Step 4: Arraylist Objarray to bind Q to Res[i] or k

Step 5: Return to users Objarray

Step 6: Display Objarray

4.6.2 Weight Calculation Algorithm using VCS

Input: Query generated from user Q, each retrieved list URL's from webpage.

Output: Each list with weight.

Here system have to find similarity of two lists: $\vec{a} = (a_1, a_2, a_3, \ldots)$ and $\vec{b} = (b_1, b_2, b_3, \ldots)$, where a_n and b_n are the components of the vector (features of the document, or values for each word of the comment) and the n is the dimension of the vectors:

Step 1: Read each row R from Data List L

Step 2: for each (Column c from R)

Step 3: Apply formula (1) on c and Q

Step 4: Score=Calc(c,Q)

Step 5: calculate relevancy score for attribute list.

Step 6: assign each Row to current weight

Step 7: Categorize all instances

Step 8: end for end procedure

4.6.3 Clustering Algorithm C-means

Input : input list of group which contains the list item LI, Train List TL, var weight

Output : Classify all the items into different clusters

Step 1: For each (item I to LI)

Step 2: For each (item j to FL)

Step 3: Define weight as double[], Hashmap <double, string>

Step 4: weight[i]=Similairty(LI[i],Fl[j])

Step 5: put into hashmap<weight[i], LI[i]Fl[j]>

End for

End for

Step 6: Sort Hashmap with desc order

Step 7: Select first value from Hasmap

Step 8: Move LI[i] to Fl[j]

4.6.4 Hash base Ranking Algorithm

Input : Hashmap <double, string>,

Output : URL list with weight

Step 1: Read each (k to Hashmap)

Step 2: evaluate each Li= $\sum_{k=0}^{n}(\text{Hashmap}[k])$

Step 3: Display Li with maximum weight

Step 4: end for

Step 5: all Li asec order

4.7 Software Requirements

1. System interfaces: Windows Operating System

2. User interfaces: User interface using Jsp and Servlet

3. Hardware interfaces

Processor :- Intel R-Core i3 2.7 or above

Memory :- 4GB or above

Other peripheral: - Printer

Hard Disk :- 500 GB

4. Software interfaces:

Front End: Jdk 1.6.0,Netbeans 7.3 or abobe

IE 6.0/above

Back-End: Mysql 5.1

4.8 Mathematical Model

The system has classified into the different sets like below

Sys={inp,process,out,analysis}

Inp= {Q1,Q2……Qn}

That is the set of input queries

List={L1,L2,L3..Ln}

$$\mathrm{LD[w]} = \sum_{k=0}^{n} \binom{n}{k[D]}$$

Extracted list from each documents

L={Wi1, Wi2, Wi3….. Win} weight of each list using below formula

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

C={c1,c2….cn} clusters of each list

[C1…Ck]= $\sum_{k=0}^{n} \mathrm{k(classify)}$

The finally system work with item or facet rank it can create the set of higher dimensions.

UrlList={URL1(w), URL2(w)…..URLn(w)}

$$UL[k] = \sum_{n=1}^{m} \text{Doc1} + \text{Doc2} \dots \dots \dots \dots \text{Docm}$$

Success condition

If(inp != Null)

Failure condition

if(UrlList == Null)

V. RESULTS AND DISCUSSIONS

We emphasis on ranking phase that considered as the main contribution of this paper. New ranking algorithm is formed to rank alike eloquent data after indexing phase. In addition to, data retrieval process become faster, easier and more accurate. The performance achieved with 99 percent relevant results in maximum time 60 ms and 1 percent only for irrelevant results. The proposed structure and ranking calculation can be additionally created for future use in distinguishing more precise semantic data from informal communities in a brief time frame. The topic of the semantic search engine has attracted large interests both from industry and research with resulting variety solutions in different tasks. There is no standardized framework that helps to monitor and stimulate the advancement in this field. In this paper, Four standard tasks of semantic search engine are discussed including crawling, indexing, ranking and finally retrieving task. This review paper shows that using SociRank produces a very unlike ranked list of news topics, which may signify that trusting only on high-frequency news topics provided by the media does not essentially give insight into what users are interested on or consider important. Taking all results into consideration emphasizes the point that MF alone is a substandard estimator of what users find interesting or consider important, and should therefore not be used in this way. SociRank, on the other hand, proves to be more capable of performing this, and so we settle that the information provided by SociRank can prove vital in commerce-based areas where the interest of users is paramount.

Method	5 d	10 d	15 d	20 d
Keyword search [3]	246	488	723	975
Form base search [4]	310	602	923	1178
Directory Search [6]	840	1520	2310	3125
SociRank [12]	180	352	533	701

Table 1: Time required in milliseconds for document retrieving to each approach

The performance analysis and efficiency calculations has visualize in below fig. 2.

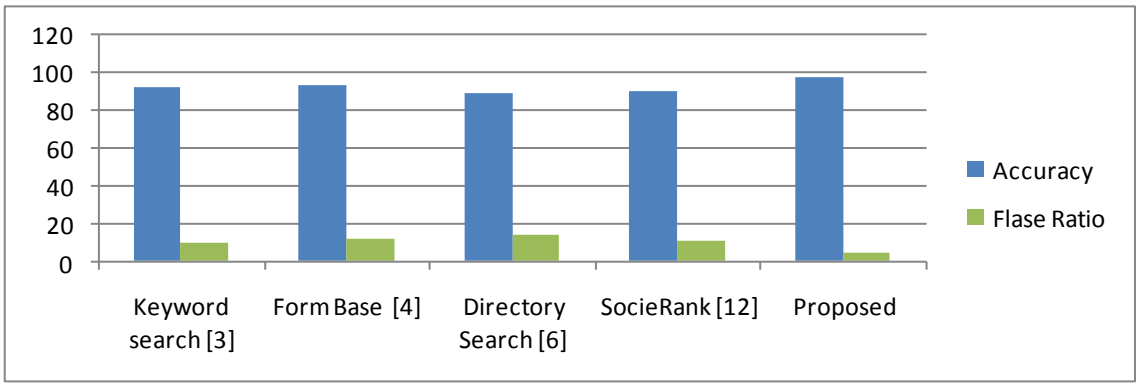

Fig.2: Proposed system efficiency calculation with existing approaches

VI. SUMMERY AND COLCLUSION

System proposed an unsupervised method SoWeRank which identifies news topics prevalent in both social media and the news media, and then ranks them by taking into account their MF, UA, and UI as relevance factors. The temporal prevalence of a particular topic in the news media is considered the MF of a topic, which gives us insight into its mass media popularity. The temporal prevalence of the topic in social media, specifically Twitter, indicates user interest, and is considered its UA. Finally, the interaction between the social media users who mention the topic indicates the strength of the community discussing it, and is considered the UI. To the best of our knowledge, no other work has attempted to employ the use of either the interests of social media users or their social relationships to aid in the ranking of topics. Our system can aid news providers by providing feedback of topics that have been discontinued by the mass media, but are still being discussed by the general population. SociRank can also be extended and adapted to other topics besides news, such as science, technology, sports, and other trends.

VII. FUTURE SCOPE

For the future environment system can focus on personalize search on user feedback sessions as well as recommendation base on user point of interest with database security is the interesting part of system.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 200

[2] T. Hofmann, “Probabilistic latent semantic analysis,” in Proc. 15th Conf. Uncertainty Artif. Intell., 1999, pp. 289–296.

[3] T. Hofmann, “Probabilistic latent semantic indexing,” in Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, Berkeley, CA, USA, 1999, pp. 50–57.

[4] C. Wartena and R. Brussee, “Topic detection by clustering keywords,” in Proc. 19th Int. Workshop Database Expert Syst. Appl. (DEXA), Turin, Italy, 2008, pp. 54–58.

[5] F. Archetti, P. Campanelli, E. Fersini, and E. Messina, "A hierarchical document clustering environment based on the induced bisecting k-means," in Proc. 7th Int. Conf. Flexible Query Answering Syst., Milan, Italy, 2006, pp. 257–269. [Online].

[6] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.

[7] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in Proc. 10th Int. Workshop Multimedia Data Min. (MDMKDD), Washington, DC, USA, 2010, Art. no. 4. [Online].

[8] W. X. Zhao et al., "Comparing Twitter and traditional media using topic models," in Advances in Information Retrieval. Heidelberg, Germany: Springer Berlin Heidelberg, 2011, pp. 338–349.

[9] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, "Finding bursty topics from microblogs," in Proc. 50th Annu. Meeting Assoc. Comput. Linguist. Long Papers, vol. 1. 2012, pp. 536–544.

[10] Derek Davis, Gerardo Figueroa, and Yi-Shin Chen, "SociRank: Identifying and Ranking Prevalent News Topics Using Social Media Factors," in 2168-2216 c 2016 IEEE.