ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

A Review on Hybrid Approach for Searching and Ranking Large Scale Web Data using Social Media Factors

Prof. Yogesh Lonkar, Dattatray Savant, Prashant Nikam, Suraj Halkude, Kamesh Patil

Department of Computer Engineering, Genba Sopanrao Moze College of Engineering, Balewadi, 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

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

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.

A novel usage based personalized page rank style algorithm was presented by Eirinaki et al [1]. This was used for ranking the web pages of the sites based on the user‟s previous navigational experiences and behaviour. Based on this algorithm a frequency selection for usage based ranking was developed. This algorithm can be mainly used to produce personalized recommendations.

The various problems in page ranking such as dangling links, historical values, false links Google Bombing, Google jacking etc were put forth by Gupta et al [2]. A concept of trust rank i.e. to place a core vote of trust on a set of seed set of reviewed pages was introduced. This trust rank can be bolted to produce more efficiency and high precision in obtaining the ranked pages.

An efficient application of content based ranking on ranking blogs was suggested by Zhu et al [3]. In this approach the set of hyperlinks are divided into browsing links, recommendation links. A diagram of blog site is created by addition of hidden link based on the content analysis. A concept of implicit and explicit ranking was introduced in this technique.

A method proposed by Feyzania et al [4] to rank the documents by constructing a connected graph of semantic documents.This graph considers only the semantic links between the documents .Thus a novel approach of content based ranking was put forth accordingly by extracting the implicit ranks .Different weights can be used to distinguish the different semantic links.

An effective approach to usage based ranking using ontology was put forth by Junfang et al [5]. The weight calculation was done using ontology tree. This weight calculation takes into account of usage information, patterns and structure. The entire ontology tree was converted into a weighted graph and then justification was applied to calculate the final weight which was then used for searching, ranking and conflict solving.

An important survey on page ranking algorithm was presented by Selvan et al[6] .The entire survey provided an efficient comparisons on these algorithms .The main focus was on the fundamental page ranking algorithms like hits and focused rank .The comparisons were drawn on the parameters of merits, demerits, performance and relevancy .

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

An effective approach of web service recommendation based on usage based ranking and QOS preferences was presented by Kang et al [7]. This algorithm is effective in classification and extraction of the results.
In [8] a novel approach in which crawler downloads only relevant documents taking advantage of migrants and thereby reducing the load on server it downloads only those web
pages that are relevant to a particular topic or a set of topics and provides information in a personalized view considering only user preferences.
In [9] Patil et al has pointed out that the inference and analysis of search goals can have a lot of advantages in improving search engine relevance and user experience by combining
web usage and web content mining.it presents a weighted technique to mine the web content catering to the user needs.
In [10] Kishorekumar et al proposed an efficient algorithm for mining the content through the extraction of search engine results and construction of dictionary for the user query and thus extraction of keywords that play a vital role in calculation of total strength of the keyword.

III. PROPOSED SYSTEM DESIGN

3.1Objectives

- 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.

3.2 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.

3.3 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).

3.4 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. The 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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

IV. RESULTS AND DISCUSSION

We focus on ranking phase that considered as the main contribution of this paper. New ranking algorithm is produced to rank similar meaningful 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 framework and ranking algorithm can be further developed for future use in detecting more accurate semantic information from social networks in a short time. 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 progress 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 different ranked list of news topics, which may signify that relying only on high-frequency news topics provided by the media does not necessarily 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 conclude that the information provided by SociRank can prove vital in commerce-based areas where the interest of users is paramount.

V. SUMMERY AND CONCLUSION

System proposed an unsupervised method SociRank 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.

REFERENCES

[1] Magdalini Eirinaki,Michalis Vazirgiannis Usage-based page rank for web personalizationin proceedings of the fifth IEEE international conference on Data mining (ICDM ″05)
[2] Samriti Gupta,Alka Jindal Contrast of Link based Web Ranking Techniques at IEEE 2015
[3] Azam Feyznia,mohsin Kahanti A link analysis based ranking method for semantic web documents at ieee proceedings of 2013
[4] Jun fang,Lei Guo, Calculation of weight of entities in ontologies by using usage based information in ieee proceedings of 2011
[5] Mercy paul Selvan,ChandraSekar ,A.Priya Darshan Survey on web page ranking algorithms in ijca (International Journal of Computer Applications) proceedings of 2012
[6] Guosheng Kang,JianxunLiu Active Web Service Recommendation Based on Usage History in ieee proceedings of service computing 2012
[7] Ashlesha Gupta,Ashutosh Dixit,AK Sharma Relevant document crawling with usage pattern and domain profile based page ranking in ieee proceedings of 2013
[8] P.Sudhakar,G.Poonkuzhal R.KishoreKumar A content based ranking for search engines in the IAENG proceedings of 2013
[9] Shital C Patil,RR keole Content and usage based ranking for enhancing search engine delivery 2014 in volume 3 issue 7International journal of Science and Research (IJSR)
[10] Safaa I. Hajeer, Rasha M. Ismail, Nagwa L. Badr, M. F. Taiba An Efficient Hybrid Usage-Based Ranking Model for Information Retrieval Systems and Web Search Engines . in ieee proceedings of 2015 6th International Conference on Information and Communication Systems (ICICS)