ISSN:0975-9646

A Novel Categorized Search Strategy using Distributional Clustering

Neenu Joseph. M[1] , Sudheep Elayidom[2]

[1]*Student, M.E ., (Computer science and Engineering) in M.G University, India,*
[2]*Associate Professor in Computer Science Department in CUSAT, India ,*

Abstract - Internet and web search has become an indispensable part of everyone's life. The usefulness of a search engine depends on the relevance of the result set it gives back. To organize and bring some order to the massive unstructured search results, search engines should group similar results together. This is a novel framework to restructure the search results by discovering different user search goals for a query. So that users with different search goals can find what they want conveniently. This is obtained by clustering the proposed feedback sessions of users constructed from user click-through logs and can efficiently reflect the information needs of users. Pseudo-documents are generated to better represent the feedback sessions for clustering. Distributional clustering is employed to cluster the pseudo-documents. Finally the search results are grouped based on the label obtained through clustering.

Index Terms – Click-through logs, feedback sessions, pseudo-documents, user search goals, distributional clustering, precision and recall

1 Introduction

In this technological era, huge amount of data is getting generated and added to the web. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the best results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve. But nowadays searching topics and detecting the accurate search results is critical and a time consuming task.

The effectiveness of information retrieval from the web largely depends on the users queries to search engines, which describes their information needs. Writing queries is never easy, because usually normal users provide queries which are short (one or two words on average) and words are ambiguous.

To make the problem even more complicated, different search engines may respond differently to the same query. Therefore, query formulation is a bottleneck issue in the usability of search engines. Typically, Web users submit a short Web query consisting of a few words to the search engines. Since these queries are short and ambiguous, how to interpret the queries in terms of a set of target categories has become a major research issue. Many different queries may refer to a single concept, while a single query may correspond to many concepts. For example, when the query “the sun" is submitted to a search engine, some users want to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun. The accurate search result that the user intends to obtain may be located in the last or second last result page. Therefore to organize this massive unstructured result set, search engines should cluster user search goals to group similar items together.

Existing works can be summarized into three classes, query classification, search result reorganization and session boundary detection. Other works focus on tagging queries with some predefined concepts to improve feature representation of queries. These methods has limitations since user feedback is not considered, many noisy search results that are not clicked by any users may be analyzed as well. All these methods only identifies whether a pair of queries belong to the same goal or mission and does not care what the goal is in detail.

The proposed work aims at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. It proposes a novel approach to infer user search goals for a query by clustering the proposed feedback sessions [1]. The feedback session is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs. Feedback sessions of multiple users are taken into account for representing different search intents. For inferring user search goals it is more efficient to analyze the feedback sessions than to analyze search results or clicked URLs directly because inside the feedback session, the clicked URLs tell what users require and unclicked URLs reflect what users do not care about.

Then, it proposes a novel optimization method to map feedback sessions to pseudo-documents[1] which can efficiently reflect user information needs. In the first step URLs are enriched with additional textual contents by extracting the titles and snippets of the returned URLs appearing in the feedback session. In the second step, an optimization method is used to combine both clicked and unclicked URLs in the feedback session. At last, clustering these pseudo-documents to infer user search goals and depict them with some keywords. Pseudo-documents are clustered by distributional clustering which is simple and effective. Finally the search results are categorized based on the user goals obtained as the label from distributional clustering.

2 LITERATURE REVIEW

2.1 Learning Query Intent from Regularized Click Graphs

Xiao Li& Ye-Yi Wang proposed this method in the year 2008; this work presents the use of click graphs in improving query intent classifiers[7], which are critical if vertical search and general purpose search services are to be offered in a unified user interface. This work investigates a completely orthogonal approach, instead of enriching feature representation; it aims at drastically increasing the amounts of training data by semi-supervised learning with click graphs. Specifically, this approach infers class memberships of unlabeled queries from those of labeled ones according to their proximities in a click graph. Moreover, it regularizes the learning with click graphs by content-based classification to avoid propagating erroneous labels.

2.2 An Implemented Rank Merging Algorithm for Meta Search Engine

Yuan Fu-yong &Wang Jin-dong proposed this method in the year 2009, which proposes an approach to improve the precision of meta search engine by using a merging method based on the combination of position and snippets/titles. It integrates two factors, the related degree between the position information of query results and the query words and the similarity between query results snippets and the query words. The results of the experiment show that, the average precision of this method is higher than popular component search engines now, and it also proved that result merging in a meta search engine[5], considering the information of the position and title/snippet of the search results at one time, the accuracy of the meta search engine will be better.

2.3 Falcons Concept Search: A Practical Search Engine for Web Ontologies

Yuzhong Qu & Gong Cheng proposed this method in the year 2011, which illustrate how the proposed mode of interaction helps users quickly find Ontologies[4] that satisfy their needs and present several supportive techniques including a new method of constructing virtual documents of concepts for keyword search, a popularity-based scheme to rank concepts and ontologies, and a way to generate query-relevant structured snippets. To facilitate concept sharing and ontology reusing, developed a Falcons Concept Search, a novel keyword based ontology search engine. The system integrates concept-level[5] search and ontology-level search by recommending ontologies and allowing filtering concepts with ontologies.

2.4 A Web Search Engine-Based Approach to Measure Semantic Similarity between Words

Danushka Bollegala & Yutaka Matsuo proposed this method in the year 2011, which proposed an empirical framework to estimate semantic similarity[3] using page counts and text snippets retrieved from a web search engine for two words. Specifically, it defines various word co-occurrence measures using page counts and integrates those with lexical patterns extracted from text snippets. To identify the numerous semantic relations that exist between two given words, also propose a novel pattern extraction algorithm[3] and a pattern clustering algorithm. The optimal combination of page counts-based co-occurrence measures and lexical pattern clusters is learned using support vector machines. The proposed method outperforms various baselines and previously proposed web-based semantic similarity measures on three benchmark data sets showing a high correlation with human ratings.

2.5 A Collaborative Decentralized Approach to Web Search

Athanasius Papagelis, Christos Zaroliagis proposed this method in the year 2012, which proposed a bottom-up approach[2] to study the web dynamics based on web-related data browsed, collected, tagged, and semi-organized by end users. This approach has been materialized into a hybrid bottom-up search engine that produces search results based solely on user provided web-related data and their sharing among users. The study shows that a bottom-up search engine starts from a core consisting of the most interesting part of the Web (according to user opinions) and incrementally (and measurably) improves its ranking, coverage, and accuracy. Finally, it also discuss how this approach can be integrated with Page Rank[2], resulting in a new page ranking algorithm that can uniquely combine link analysis with users preferences.

3 SYSTEM ARCHITECTURE

This work aims at categorizing or grouping the search results of an ambiguous query using labels represented by keywords. It involves discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. It proposes a novel approach to infer user search goals for a query by clustering the proposed feedback sessions. The feedback session is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs. Feedback sessions of multiple users are taken into account for representing different search intents. Then, it proposes a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. Finally, pseudo-documents are grouped using distributional clustering to infer user search goals and search results are restructured with these keywords.

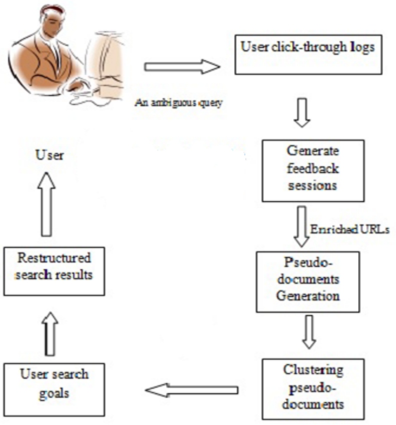
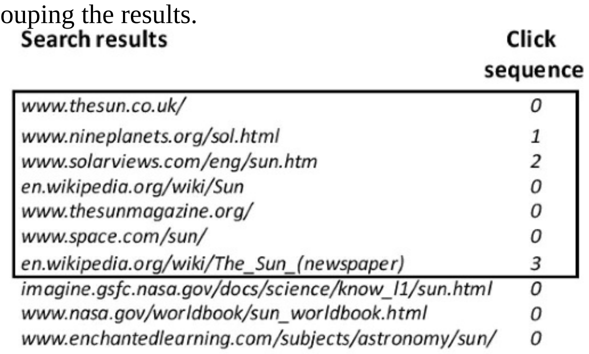

User click-through logs
An ambiguous query
Generate feedback sessions
Enriched URLs
Pseudo-documents Generation
Clustering pseudo-documents
User search goals
Restructured search results
User

Figure 1: Architecture of result restructuring framework

This work has three major modules:
1. Generation of feedback sessions
2. Pseudo-Document Formation
3. Clustering of Pseudo-documents

Generation of feedback sessions

Sessions for a web search include series of activities of a single user between the login and logout time period to satisfy a single information need. It consists of some clicked and unclicked search results. Here, feedback session consists of both clicked and unclicked URLs and ends with the last URL that was clicked in a single session. In this approach, feedback sessions of multiple users are taken into account for obtaining different search intents for grouping the results.



Search results	Click sequence
www.thesun.co.uk/	0
www.nineplanets.org/sol.html	1
www.solarviews.com/eng/sun.htm	2
en.wikipedia.org/wiki/Sun	0
www.thesunmagazine.org/	0
www.space.com/sun/	0
en.wikipedia.org/wiki/The_Sun_(newspaper)	3
imagine.gsfc.nasa.gov/docs/science/know_l1/sun.html	0
www.nasa.gov/worldbook/sun_worldbook.html	0
www.enchantedlearning.com/subjects/astronomy/sun/	0

Figure 2: A single feedback session

Pseudo-Document Formation

Pseudo-document formation mainly deals with representing the URLs in the feedback sessions with some keywords. Keywords are obtained from enriched URLs which is a combination of titles and snippets of corresponding URLs. The titles and snippets of URLs are extracted in an offline manner and are combined together to form a keyword set. Pseudo-documents are generated from these enriched URLs by considering the keywords of only clicked URLs from the feedback sessions.

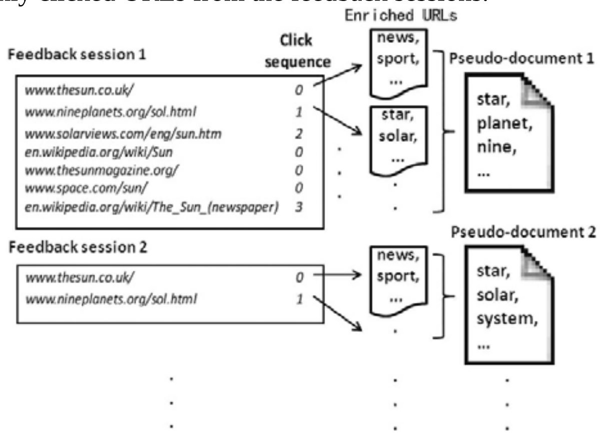


Figure 3: Mapping feedback sessions to pseudo-documents

Clustering of Pseudo-documents

Pseudo-documents are grouped together using distributional clustering as shown in fig: 4. Clustering based techniques involve partitioning the data into groups which contains similar objects. It is used to improve the efficiency of the result by making groups of data. The goal of a clustering algorithm is to group objects into meaningful subclasses. Clustering can be used to generate class labels for a group of data. For that a dictionary is generated which consist of keywords extracted from URLs. It is compared with the keywords extracted from feedback sessions of users. Finally, it forms different clusters according to the search intent of users and returns label for restructuring results.

```
Algorithm 1 Distributional Clustering
Input: Keyword list
generate the dictionary of keywords for classification.
repeat
    for each data point x∈D do
        compare it with the dictionary elements
        assign x to the clusters// each cluster represented by a label
    end for
    merge the clusters containing similar elements
until the stopping criterion is met
Output: set of clusters with label
```

Distributional clustering is an unsupervised dimensionality reduction technique that has high document classification accuracy. This approach clusters words into different groups based on the distribution of class labels associated with it. The key benefits of document clustering are higher classification accuracy and smaller classification models. Thus it reduces the number of redundant as well as noisy features.

In this framework, distributional clustering works by generating a dictionary of keywords which is obtained from the title+snippet combination extracted from the URLs. It is then compared with the keyword set obtained from the feedback sessions of users. For example, if one user searches for sun star and the other for sun newspaper, keywords obtained from their feedback sessions contains words corresponding to both search intentions. This keyword set is compared with the dictionary and the words corresponding to sun star is grouped to one cluster and sun newspaper to another cluster. Each cluster is represented by their corresponding labels.

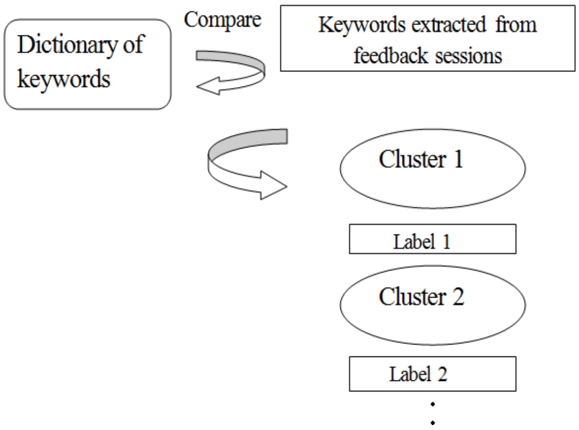
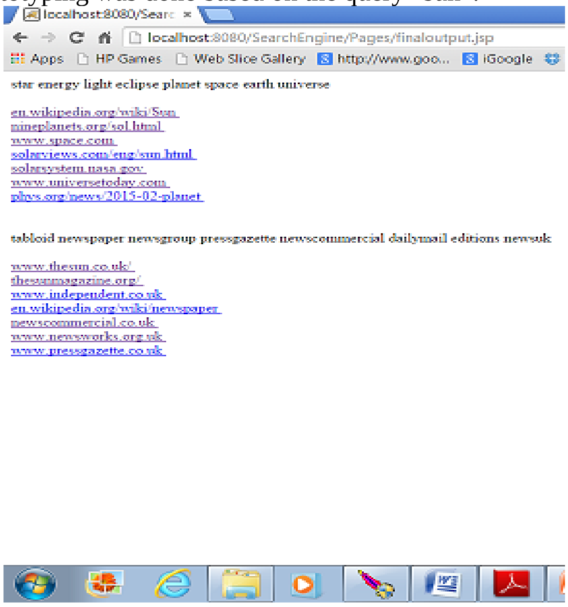


Figure 4: Distributional clustering

Labels are obtained by performing dimensionality reduction of cluster keywords. Finally the original url set used for simulating the system is reorganized or categorized based on these labels.

4 Experimental Results

The proposed work was simulated by creating a search engine using the dataset from Google search engine. The prototyping was done based on the query "sun".



star energy light eclipse planet space earth universe

en.wikipedia.org/wiki/Sun
nineplanets.org/sol.html
www.space.com
solarviews.com/eng/sun.html
solarsystem.nasa.gov
www.universetoday.com
phys.org/news/2015-02-planet

tabloid newspaper newsgroup pressgazette newscommercial dailymail editions newsuk

www.thesun.co.uk/
thesunmagazine.org/
www.independent.co.uk
en.wikipedia.org/wiki/newspaper
newscommercial.co.uk
www.newsworks.org.uk
www.pressgazette.co.uk

Figure 5: Restructured search results

The URLs were selected and stored in the database for performing the preprocessing steps. This work was done assuming two users with different search intentions. One user selected URLs relating to the natural knowledge of sun and the other opted for "sun" newspaper in United Kingdom. Feedback sessions were generated from both the users. A Feedback session consisted of clicked and unclicked URLs and ended with the last URL in a single session. Likewise, all the feedback sessions since the last click of both the users were taken for further processing. By considering feedback sessions from multiple users the reliability and effectiveness of search were increased. From the URLs of feedback session, snippets and titles were extracted and combined together using an optimization criterion. This enriched URLS were converted to pseudo-documents by considering the title+snippet combination of only clicked URLs in the feedback session. These pseudo-documents were clustered based on distributional clustering. The output of distributional clustering i.e. labels were taken to the next stage of processing. Labels were actually obtained by the dimensionality reduction of cluster elements. As shown in fig: 5, the URLs that correspond to the query "sun" were finally categorized or organized using these labels.

Performance Evaluation

In an information retrieval scenario, the instances are documents and the task is to return a set of relevant documents given a search term. Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labelled as belonging to the positive class) divided by the total number of elements labelled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labelled as belonging to the class).

Recall in this context is the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labelled as belonging to the positive class but should have been). This system considers relevant documents as the URLs that correspond to the search goal out of the retrieved documents. The result set obtained from Google was compared with that of the new system with the same query.

$$\text{Precision} = \frac{(\text{relevant docs}) \cap (\text{retrieved docs})}{\text{retrieved docs}}$$

$$\text{Recall} = \frac{(\text{relevant docs}) \cap (\text{retrieved docs})}{\text{relevant docs}}$$

Precision	Recall
$P_g = 5/10 = 0.5$	$R_g = 5/9 = 0.55$
$Po = 7/7 = 1$	$R_o = 7/7 = 1$

5 Conclusion

Due to the enormous amount of information available in the internet, improving the usability and applicability of search engines has become an important area of research. This work proposes a novel method for restructuring the search results considering the user behavior. The user behavior is analyzed by considering the click-through logs and generating feedback sessions. Initially, the feedback sessions of multiple users are analyzed to infer user search goals rather than search results or clicked URLs directly. Both the clicked URLs and the unclicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions. Therefore, feedback sessions can reflect user information needs more efficiently. Secondly feedback sessions are mapped to pseudo-documents to approximate goal texts in user minds. The pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets. Based on these pseudo-documents, user search goals can then be discovered and depicted with some keywords. Pseudo-documents are clustered using distributional clustering. Finally, the URLs were categorized based on the labels obtained from clustering. In reality, this approach can discover user search goals for some popular queries when users submit one of the queries, the search engine can return the results that are categorized into different groups according to user search goals. Thus, users can find what they want conveniently.

ACKNOWLEDGEMENT

I express my sincere thanks to all respected personalities who had guided, inspired and helped me in the successful completion of this paper. I express my thanks to The Lord Almighty for guiding me in this endeavor and making it a success. I am extremely grateful to my guide Dr. Sudheep Elayidom, Associate Professor, Department of Computer Science & Engineering, CUSAT, Ernakulam, for the inspiring and sincere guidance.

REFERENCES

[1] A New Algorithm for Inferring User Search Goals with Feedback Sessions Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, March,2013

[2] A Collaborative Decentralized Approach to Web Search by Athanasios Papagelis & Christos Zaroliagis, IEEE 2012

[3] A Web Search Engine-Based Approach to Measure Semantic Similarity between Words by Danushka Bollegala & Yutaka Matsuo, IEEE 2011

[4] Falcons Concept Search: A Practical Search Engine for Web Ontologies by Yuzhong Qu & Gong Cheng, IEEE 2011

[5] An Implemented Rank Merging Algorithm for Meta Search Engine by Yuan Fu-yong & Wang Jin-dong, IEEE 2009

[6] Context-Aware Query Suggestion by Mining Click-Through and Session Data by Huanhuan Cao & Daxin Jiang, ACM 2008

[7] Learning Query Intent from Regularized Click Graphs by X. Li, Y.-Y Wang, and A. Acero, ACM 2008

[8] Building Bridges for Web Query Classification by Dou Shen & Jian-Tao, ACM 2006

[9] What You Seek is What You Get: Extraction of Class Attributes from Query Logs by Marius Pasca & Benjamin Van Durme, IJCAI 2007

[10] Learning to Cluster Web Search Results by Hua-Jun Zeng & Qi-Cai, ACM 2004