Opinion Ranking based on Lists in Search Engines

Waleed A. Almutairi
Computer Research Institute.
King Abdulaziz City for Science and Technology (KACST)
Riyadh, Saudi Arabia
Walmutairi@kacst.edu.sa

Abstract—**Search engines have improved gradually for the last few years. The Artificial Intelligence (AI) field has played a big role in this area. The Internet has a lot of raw and informative data. Most popular search engines try to identify and understand the connection between those data and make it useful.**

In this paper, I propose an AI technique to classify and add some ranking and scoring to the listing phrases across the documents that have the same key-phrase. If the document has a listing like 1, 2, 3… or A, B, C… or I, II, III, or bullet points and so on. The inverted index will have the keys with their lists plus their scores as a phrase. When we search for a key that has a listing in it, the query will search for the key for matches through the index. Then we will get the results organized by their Ranking. *(Abstract)*

Keywords-phrases Information Extraction; AI; Listing; Ordering; Classification; Pattern recognition; Named Entity (key words)

I. INTRODUCTION

Data are available via the Internet, and it is easy to get. But how to extract the information and get some knowledge about something from this data, and this is the tricky part? Text classification and Named Entity recognition (NER) is useful to get some sense about the data, for example like the Question Answering (QA) algorithms is famous and widely used in the Internet. The QA algorithm tries to recognize and extract the Questions and their answers from the data. Then add those data to the index, and then when the query met the question the accurate answers will be retrieved. A simple question answering systems approaches that are using limited background knowledge can improve the accuracy of the results [6].

Data mining or in our case its text mining, it's to analyze the text, look for a pattern and study the words frequencies. This field can give us a glance of what this document is talking about by calculating many aspects. Many systems have invested some money and time to extract some useful output, so that they can use. In addition, search engines uses text mining to increase the accuracy of their results.

In this paper I propose a way to rank the listing or ordered sentences. This list is ordered by users' opinion. And by gathering the documents across the Internet, we may have two or more opinions about one thing. For example, we may have people listing what they prefer better like, restaurants, places, cities, and people or can be something else.

II. RELATED WORK

The subject of name entity and pattern recognition have improved gradually in the last decades, Moreover, Text classification have been a hot topic to many researchers to extract some knowledge from the data.

In the area of Question Answering System, an experiment was made of Vietnamese Question Answering system by combining Snowball system and semantic relation extraction using search engine. The result was 89.7 %precision and 91.4 recall [1]. On other Question answering level, there has been development on designing an Automated Question Generation from extracting sentences within a document as a source of question/answer data [2]. Muthukrishanan Umamehaswari et al [3] have improved question answering system by semantic reformulation the idea here is to generate the pattern from the web based on their lexical semantic and syntactic constrain. These constrains are defined in the question answering system to evaluate and rank the candidate answer. Eric Sneiders [4] who designed a two layer system FAQ model and a site search engine to capture reoccurring queries , also natural language questions, and present semantically matching answers, while the search engine will work as a statistical keywords matching .In An specialized question answering system Zhang Wei et al [5] Design an Influenza question answering system based on multi-strategies. The system will run on a multi-level querying. first, the user input the question. And if the question was found, then show the result. Otherwise, the influenza information ontology will utilize and a shallow semantic analysis to generate a question vector based on the user's question. Liu xiaoli [7] proposed semantic software architecture for pattern based user interactive question answering system. The system has three dimensions, question customizing, question analysis and question refinement.

In the area of the index and the data in the inverted index, the index is the most important part in search engines and in searches in general. Feng Yu et al [8] has combined CRF model with document structure. In addition, the key-phrases not only represent the main content of the document, but also reflect the specialty of this document. In this method the result shows an improvement in precision and recall rate. Qiuying Bai et al [9] has proposed a model for a new inverted index, which is a combination inverted index (CII),

doi:10.4304/jait.5.3.71-73

and contains the prime inverted index, appendix inverted index and deleted file list. As a result this model increased the retrieval efficiency and kept the recall rate as the same.

III. Problem Definition

We have a lot of data that presented in lists. The Internet has millions of blogs and documents that have an opinion about something; like the causes of an action, or an opinion about something. For example, the questions "What are the best java books", "What are the top schools in the USA?", "What are the causes of global warming?" and "What are the best Italian restaurant in New York?" Suppose we have those data in users' blogs or documents. And it was represented as a list and ordered by number or any kind of listing scheme.

To solve those kinds of questions we need to identify this pattern. Also, we want to increase the frequented answers to have more weight, so then when we asked about something the result will be sorted by its weight.

A. Pattern Recognition

To recognize the lists and the ordered sentences there are a pattern to look for. Like these patterns for example here are some of these patterns:

XXXXXXXXXXX

1. ……..
2. ……..
3. ……..

XXXXXXXXXXX

I. ……
II.…….
III. ……..

XXXXXXXXXXX

- ……
- ……
- ……

This lists and their order is based on the writer opinion. And it is will be different from person to another, based on what he/she likes. Also the order may change based on the opinion. If the classifier recognize this forms then. Go to the next step.

B. scheme and index

The lists and the key-phrases will be added to the index, but with the reference to its location, the Order number in the paragraph, will have a higher weight. If it appears again in other documents or pages then the score will be increased. The parser will check if the key-phrase appeared more than once or not, if yes then do some work. We will see the algorithm in the next section.

In Figuer1 we have three documents and we assumed all the three subjects are the same, but the order of them are different.

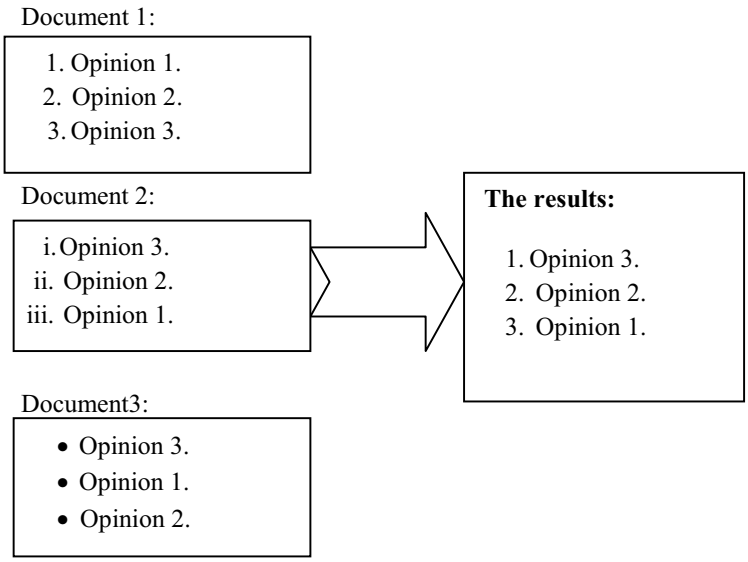


Figure1. Results based on their frequencies. All the three documents have the same Key-phrase

IV. Algorithms

In this section, we will describe the algorithm after we crawled and fetched documents and web pages. That means from parsing data, the representation of the proposed pattern in the index, and the algorithm to search for a key-phrase by the users:

A. The parser and index algorithm:

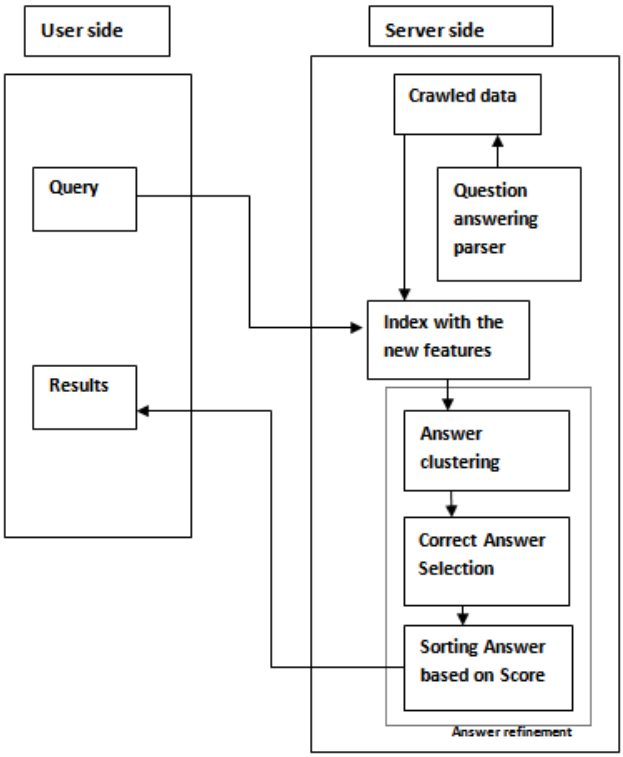
- Run_classifier(); to identify the proposed scheme.
- Declare a new field in the index that will hold the new key-phrases <ListOrder>.
- This field will hold the weight<Weight> and the location of those phrases in the paragraph<Order>.
- Check for similarity() If same key-phrases or the key appeared again, increase the <Weight> and add a reference to the old list<Order>.
- After finishing parsing the data add the <Weight> and resort the <Order> based on their frequencies. Like in figure1.
- Add this new field to the index.

B. Searching for a key-phrase:

- If for example we have the query: “What are the top ranked schools in the USA?”
- Remove_Stop_Words();
- Convert_To_small_letters();
- Run_Stemmer();
- Look for the keys” top” “ranked “school” and “usa” and have <Order>; in the index.
- Show_result (Get.order ()); the new rank in the index that was computed in the end.

V. System Architecture

In this section, we have the Architecture with small changes from the usual index architecture design, in parsing the data after crawling it. The system will have two-sides user side and server side, then after parsing the data there will be new features in the index that indicate the scores for each answers for the questions.



VI. Conclusion

In the paper, we proposed a novel way to search across the lists in data. In detecting then adding those lists to the index would be a new information added to your index and to your search engine, to give a glance of what people think or what their opinion about something.

In future work, we will implement the proposed method with sample data to show the statistics and see how useful it is.

References

[1] Vu Mai Tran; Vinh Duc Nguyen; Oanh Thi Tran; Uyen Thu Thi Pham; Thuy Quang Ha, "An Experimental Study of Vietnamese Question Answering System," *Asian Language Processing, 2009. IALP '09. International Conference on* , vol., no., pp.152,155, 7-9 Dec. 2009

[2] Han-joon Kim; Han-Joon Kim, "Design of Question Answering System with Automated Question Generation," *Networked Computing and Advanced Information Management, 2008. NCM '08. Fourth International Conference on* , vol.2, no., pp.365,368, 2-4 Sept. 2008

[3] Umamehaswari, M.; Ramprasath, M.; Hariharan, S., "Improved Question Answering System by semantic refomulation," *Advanced Computing (ICoAC), 2012 Fourth International Conference on* , vol., no., pp.1,4, 13-15 Dec. 2012

[4] Sneiders, E., "Automated FAQ answering with question-specific knowledge representation for web self-service," *Human System Interactions, 2009. HSI '09. 2nd Conference on* , vol., no., pp.298,305, 21-23 May 2009

[5] Zhang Wei; Zhang Xuan; Chen Junjie, "Design and implementation of influenza Question Answering System based on multi-strategies," *Computer Science and Automation Engineering (CSAE), 2012 IEEE International Conference on* , vol.1, no., pp.720,723, 25-27 May 2012

[6] McGuinness, D.L., "Question answering on the semantic Web," *Intelligent Systems, IEEE* , vol.19, no.1, pp.82,85, Jan-Feb 2004

[7] Liu xiaoli; Wu Guoqing; Jiang Min; Yang Min; Wang Weiming, "Software architecture for a pattern based Question Answering system," *Software Engineering Research, Management & Applications, 2007. SERA 2007. 5th ACIS International Conference on* , vol., no., pp.331,336, 20-22 Aug. 2007

[8] Feng Yu; Hong-Wei Xuan; De-quan Zheng, "Key-Phrase Extraction Based on a Combination of CRF Model with Document Structure," *Computational Intelligence and Security (CIS), 2012 Eighth International Conference on* , vol., no., pp.406,410, 17-18 Nov. 2012

[9] Qiuying Bai; Chi Ma; Xuechang Chen, "A new index model based on inverted index," *Software Engineering and Service Science (ICSESS), 2012 IEEE 3rd International Conference on* , vol., no., pp.157,160, 22-24 June 2012