ISSN (Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

Relation-based Wikipedia Search System for Factoid Questions Answering

Adam Brzeski, Tomasz Boiński

Department of Computer Architecture, Faculty of Electronics, Telecommunications and Informatics,
Gdańsk University of Technology, Poland

ABSTRACT: In this paper we propose an alternative keyword search mechanism for Wikipedia, designed as a prototype solution towards factoid questions answering. The method considers relations between articles for finding the best matching article. Unlike the standard Wikipedia search engine and also Google engine, which search the articles content independently, requiring the entire query to be satisfied by a single article, the proposed system is intended to solve queries by employing information contained in multiple articles. Although still a keyword search, the method can be further employed in natural language questions answering, when accompanied with a question processing module. The method assumes that queries are formulated in a form of a list of Wikipedia articles. The possible solutions are then evaluated, however not by attempting to understand the meaning of the text, but by a simple method of estimating the distance between articles by measuring articles' references or appearances in other articles, leading finally to returning a single article as an answer for the query.

KEYWORDS: natural language processing, search engine, semi-structured text, open-domain questions

I. INTRODUCTION

The transition from keyword-based search to natural language questions answering has been long awaited, and even though more and more often it is announced to be finally achieved, it is in fact still unavailable. While the modern natural language solutions are trying to gain trust and become popular, keyword search dominates most of the querying systems. It also applies to Wikipedia, which in the meantime has become a huge and powerful base of knowledge. Wikipedia is therefore a great source of information for question answering systems, and thereby searching Wikipedia for particular information becomes highly interesting and important task. Efficient search engine is obviously also important for regular users.

At the same time, the standard Wikipedia keyword search engine remains a solid but simple solution. And the search is in fact an non-trivial problem. Although Wikipedia articles in many case have a form of unstructured text, there is also well structured content of infoboxes that can be extensively utilized. Also, Wikipedia contains a great net of links between articles. All of this information should be considered by an efficient search engine. The Wikipedia search systems remains simple in a sense that it attempts to match the articles independently to the entire query at once. The structure of Wikipedia, however, opens a possibility of finding matching articles not only basing on their own content, but also basing on the content of articles closely related to them. In that sense, a query can be satisfied by a group of related articles. In the presented method we attempt to follow this approach.

II. RELATED WORK

Other approaches for alternative Wikipedia searches were already proposed. Some of them introduce new techniques for querying Wikipedia. Hahn et al. [1] presented a faceted search system based on DBpedia. Yan et al. [2] also used faceted approach in order to provide better browsing of search results. Hu et al. proposed a search results re-ranking system considering the quality of articles [3]. Boiński and Brzeski presented a facts extraction method for Polish Wikipieda [4]. Szymański, in turn, presented systems for Wikipedia search results clustering [5] and articles categorization [6]. Wikipedia is also widely utilized as a knowledge base for question answering systems. Recently, Ryu et al. presented a thorough approach for question answering employing article infoboxes, content, structure, category

ISSN (Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

structure and definitions [7]. Interesting work was also presented by Chu-Carroll and Fan, who utilized Wikipedia data for generating candidate answers for open-domain questions [8].

III. THE PROPOSED METHOD

The proposed method assumes that the query is given in a form of a set of Wikipedia articles' titles. The task is to find the most relevant Wikipedia article, which best connects the given phrases and therefore is the expected result of the query. The proposed algorithm can be briefly formulated in the following way:

Input: query consisting of one or more phrases corresponding to Wikipedia articles
Output: answer (a Wikipedia article)
Steps:

1) Establish a set of candidate solutions
2) Evaluate candidate solutions with relevance function f
3) Return the solution with the highest score

In the first step the candidate solutions set is established by performing a sequence of regular Wikipedia searches using particular phrases of the query. For each of the phrases the first search result is selected as the phrase article. Then, for each of the phrase articles, a list of all linked articles is acquired and appended to the candidate solutions set. In that point it is assumed that none of the phrases itself is the answer to the query, so the phrase articles are not included into the set. Typically the candidate solution set consists of a few hundred to few thousand articles, depending on the query length and the phrases used. In the second step candidate articles relevance is evaluated using function f defined by equation 1.

$$f(article, Query) = \sum\nolimits_{phrase_i \in Query} g(article, phrase_i) \tag{1}$$

$$g(article, phrase) = \sum_{i=1}^{n} R_i \cdot w_i + match\_bonus \cdot w_m \tag{2}$$

Where:

$article$, $phrase$: Wikipedia articles

$$R_i = \begin{cases} 2, & \text{if there is relation } i \text{ between } article \text{ and } phrase \text{ in both directions,} \\ 1, & \text{if there is relation } i \text{ between } article \text{ and } phrase \text{ in one direction,} \\ 0, & \text{otherwise} \end{cases}$$

$$match\_bonus = \begin{cases} 1, & \text{if } \prod_{i=1}^{n} R_i > 0 \\ 0, & \text{otherwise} \end{cases}$$

w_i: relation weights
w_m: match bonus weight

Three types of potential relations between the articles are defined. The descriptions and conditions of the relations are presented below. Finally, for the evaluation of f function the following weight values were assumed: $w_1 = 10$, $w_2 = 5$, $w_3 = 1$ and $w_m = 20$.

Rel 1 A close connection between articles. Article A is related to article B, if article B title is equal to the type of infobox used in article A, or if it appears in the first sentence of article A. If the first sentence is long, only

ISSN (Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

the first part of it is considered, by trimming it on the second comma character from the beginning of the sentence.

Rel 2 (An important reference of article) Article A is related to article B, if article B appears in article A's infobox.

Rel 3 (Link to article) Article A is related to article B, if article A links article B

IV. RESULTS

The experiments were carried out on a set of 20 queries extracted from one of the episodes of a popular Polish general knowledge quiz show. The natural language questions were transformed into queries consisting of valid phrases, that is phrases leading to particular Wikipedia articles. However, not all of the quiz questions could be utilized in this manner. The questions converted into the keywords must remain clear and answerable. Therefore questions, in which a verb plays key role were eliminated, since verbs can hardly be represented by Wikipedia articles. Also any "yes-no" questions had to be excluded, as well as "choose one" questions, questions with numerical answers, and questions not covered by English Wikipedia.

The set of 20 queries was obtained from a random subset of the valid questions. Ten of the queries were used as a training set for developing and tuning the method, while the remaining queries were used as a test set. All considered queries were presented in table I.

In order to provide a reference for the proposed method, we also test the performance of Google search and the regular Wikipedia search on the queries by checking the first returned article. For optimizing the reference searches, various combinations of double quotes in the queries were tested in order to obtain the best result. Also, similarly as in the proposed method, the search results which corresponded to the query phrases were ignored, in order to increase the chance of the correct answer appearing on the top. The results are presented in table II.

Id	Query	Expected answer
	Training set	
1	"capital city" "San Marino"	City of San Marino
2	"Robert Peary" "Geographical pole" "1909"	North Pole
3	"country" "Lorraine region"	France
4	"chemical element" "CU"	Copper
5	"country" "Confucius"	China
6	"peninsula" "Bay of Puck"	Hel Peninsula
7	"language" "Argentina"	Spanish language
8	"river" "Poland" "Ukraine" "Belarus"	Bug River
9	"priest" "Trojan Horse"	Laocoön
10	"actor" "The Fugitive 1993" "Richard Kimble"	Harrison Ford
	Test set	
11	"country" "People's Party for Freedom and Democracy" "Party for Freedom"	Netherlands
12	"musical instrument" "Nicanor Zabaleta"	Harp
13	"novel" "trilogy" "Henryk Sienkiewicz" "Jasna Góra"	The Deluge (novel)
14	"country" "Tyrol state"	Austria
15	"codename" "Polish Scouting Association" "underground"	Gray Ranks
16	"planet" "Callisto" "Europa moon"	Jupiter
17	"dynasty" "Władysław I Herman"	Piast dynasty
18	"character" "Winnie-the-Pooh" "donkey"	Eeyore
19	"character" "Round Table" "traitor"	Mordred
20	"mountain range" "Black Sea" "Caspian Sea"	Caucasus Mountains

Table I: The queries from the traning and test sets along with expected answers.

ISSN (Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

Id	Answers		
	Google	**Wikipedia**	**Proposed method**
		Training set	
1	✓	✓	✓
2	Matthew Henson	Roald Amundsen	✓
3	Alsace-Lorraine	Metz	✓
4	✓	✓	✓
5	Confucianism	Confucius Institute	✓
6	Gdańsk Bay	✓	✓
7	Languages of Argentina	Languages of Argentina	Buenos Aires Herald
8	✓	✓	Vistula
9	✓	Trojan War	✓
10	The Fugitive (TV series)	The Fugitive (TV series)	✓
		Test set	
11	Geert Wilders	Liberalism in the Netherlands	✓
12	Joaquín Rodrigo	Marat Bisengaliev	✓
13	✓	✓	Aleksandra Billewiczówna
14	Tyrol	North Tyrol	✓
15	✓	✓	✓
16	Galilean moons	Galilean moons	✓
17	Władysław I the Elbow-high	Bolesław III Wrymouth	✓
18	✓	✓	✓
19	Guinevere	✓	✓
20	✓	✓	✓

Table II: The results acquired from Google search, regular Wikipedia search and the proposed method. Checkmarks denote correct answers.

V. Discussion

In the first place, the experiments showed that about half of the queries could be correctly resolved by the standard Wikipedia search engine. A similar result was achieved by Google search. The two search systems resolved the queries, which were entirely contained in the proper article text. This leads to a conclusion, that large fraction of the quiz questions could be answered basing on a single, but appropriate article. Interestingly, in 2 cases out of 3, where the proposed method failed to return the correct answer, both of the standards search systems succeeded. The reason was that the proposed method doesn't yet consider the actual article text, except the first sentence and links, while the standards search focused on the text and therefore managed to spot the correct article. Furthermore, in some cases the method mistakenly overvalued relations of other articles, bringing it to the top of the results list. Except full article content consideration, Wikipedia and Google searches also outperformed the proposed method by accepting multiple forms of words or named entities. The proposed method also still lacks consideration of synonyms, words derivations and Wikipedia redirections, which leads to somehow strong constraints on the queries, which limits Google and Wikipedia search systems in a lower degree. The proposed method, in turn, significantly outperformed the other search systems in terms of collecting information from multiple articles. While the standard search systems failed to answer all of the queries of that kind, the proposed method failed only once. This is quite a satisfactory and important result, especially that, again, about half of the queries required combining the information from multiple articles in order to be answered.

VI. Conclusions

The presented prototype method achieved promising results in solving the posed problem of a keyword search in Wikipedia articles base. The correct answers rate turned to be almost twice higher than for the regular Wikipedia search

ISSN (Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

or Google search. The results on detecting cross-article knowledge may suggest a potential for use in natural language question answering. The test set used, however, was rather small and the experiment requires confirmation on a larger set. Further refinements are also expected, mainly by considering the full text of an article and introducing isA-like relation, possibly by employing DBpedia or Yago ontologies.

REFERENCES

1. Rasmus Hahn, Christian Bizer, Christopher Sahnwaldt, Christian Herta, Scott Robinson, Michaela Bürgle, Holger Düwiger and Ulrich Scheel, 'Faceted Wikipedia Search'. *Business Information Systems, Lecture Notes in Business Information Processing*, Vol. 47, pp. 1–11, 2010.
2. Ning Yan, Chengkai Li, Senjuti B. Roy, Rakesh Ramegowda and Gautam Das, 'Facetedpedia: enabling query-dependent faceted search for wikipedia', *In Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)*, pp. 1927–8, 2010.
3. Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw and Ba-Quy Vuong, 'On improving wikipedia search using article quality', *In Proceedings of the 9th annual ACM international workshop on Web information and data management (WIDM '07)*, pp. 145–52, 2007.
4. Tomasz Boiński and Adam Brzeski, 'Towards Facts Extraction from Texts in the Polish Language', *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, No. 8, 2014.
5. Julian Szymański, 'Self–Organizing Map Representation for Clustering Wikipedia Search Results', *Intelligent Information and Database Systems, Lecture Notes in Computer Science*, pp. 140–9, 2011.
6. Julian Szymański, 'Wikipedia Articles Representation with Matrix'u', *Distributed Computing and Internet Technology, Lecture Notes in Computer Science*, pp. 500–10, 2013.
7. Pum-Mo Ryu, Myung-Gil Jang and Hyun-Ki Kim, 'Open domain question answering using Wikipedia-based knowledge model', *Information Processing and Management*, Vol. 50, No. 5, pp. 683–92, 2014.
8. Jennifer Chu-Carroll and James Fan, 'Leveraging Wikipedia Characteristics for Search and Candidate Generation in Question Answering', *In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.