TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 8, August 2014, pp. 6380 ~ 6385
DOI: 10.11591/telkomnika.v12i8.5143

Conceptual Search Based on Semantic Relatedness

Abdoulahi Boubacar[1], Zhendong Niu[2]
Beijing Institute of Technology, School of Computer Science
*Corresponding author, e-mail: abd_boubacar@yahoo.com[1], zniu@bit.edu.cn[2]

Abstract

Traditional search engines based on syntactic search are unable to solve key issues like synonymy and polysemy. Solving these issues leads to the invention of the semantic web. The semantic search engines indeed overcome these issues. Nowadays the most important part of the data remains unstructured documents. It is consequently very time consuming to annotate such big data. Concept based retrieval systems intend to manage directly unstructured documents. Semantic relationships are their main feature to extend syntactic search. In most of the methods implemented so far, concepts are used for both indexing and searching. Words remain the smallest unit to process semantic relatedness. The differences persist in the way that concepts are represented, mapped to each other, and managed for the sake of indexing and/or searching. Our approach is based on Wikipedia concepts. Concepts are represented as an undirected graph. Their semantic relatedness is computed with a distance derived from a semantic similarity measure. The same distance is used to calculate both semantic relatedness and query matching.

Keywords: c*oncept analysis, information retrieval, semantic relatedness*

1. Introduction

To implement a concept based retrieval system, the first question is always “what is a concept”. There are many answers to this question. A concept may be any idea or thing that has a meaning by itself. Some concepts are mono-word while others are multi-word. A concept can be represented by a word, a sentence fragment, a whole sentence or an entire document. Concepts has been defined as WordNet entrees [1, 2]. The WordNet approach has solved certainly the synonymy problem. Query can be expanded using the synonyms. To solve polysemy problem the semantic web search engines use ontologies. The method is perfect in term of precision [3-5]. Another approach is based on word’s frequencies according to a given corpus. The Latent Semantic Analysis (LSA) [6, 7], presents a reduction method that optimizes concept extraction for a large scale of corpus. The LSA method uses matrix factorization instead of human comprehensible knowledge. Our approach is based on Wikipedia articles. Each of the selected articles represents one concept. Incomplete articles are not selected. The second issue to deal with is the choice of the tool. Tools could be statistical, probabilistic etc. We have chosen to use only one tool: the semantic distance between the three different entities that are queries, concepts and documents. The semantic distance is used to build an undirected graph of concepts. We consider that each concept may have a link to other concepts. We did not group the concepts into partitions. For this reason the graph representation seems to be the most adequate. Opposite the methods based on Formal Concept Analysis [8] and [9], we did not establish a hierarchy between concepts.

2. Related Work

The best choice for indexing is still unclear in information retrieval. Words or concepts, which one is the better? Yiming Yang [10] and Hersh et al [11] have investigated the best way to represent a document. For a sake of performance, indexing with words as lexical units is better than indexing with concepts. For a sake of relevance, indexing with concepts as semantic units is better than indexing with words. In a concept based retrieval system any idea, person, thing etc. can be a concept [12]. In such system users do not need to find a magic word that can connect them to the information they seek. William A. Woods [13] is one of the researchers who

Received November 14, 2013; Revised April 3, 2014; Accepted April 17, 2014

developed very early (1997) a conceptual indexing method based on taxonomy where concepts are presented at sentence level. His method, does not use a hierarchy of concepts in contrast with Wright et al [14] and Chen et al [15]. Hierarchical relationships have been used by Hersh et al to implement SAPHIRE. SAPHIRE [16] combined both semantic and probabilistic methods to develop a heuristic retrieval environment. Concept based systems have been developed as an alternative to syntactic search [17] placing words into a context [18]. Most of the models developed to overcome issues related to syntactic search are not language dependent [19]. Concept can be extracted from query [20] or from documents [21]. Comparison have been made by Dobsa and Basie [22] between Latent Semantic Indexing and concept based indexing in information retrieval. Their results have shown that concept indexing is computationally more efficient than Latent Semantic Indexing. Different concept based web applications have been built using concept recognition [23, 24] for query answering. A survey conducted by Haav and Lubi [25], through thirty six concept based information retrieval tools on the web, have shown the need of improvement in different directions. Our approach is based on semantic relatedness. The question we intend to solve is how to efficiently use the concepts semantic relatedness to improve the state-of-the-art methods. For that reason we need an appropriate semantic distance and a pertinent concept representation.

3. Semantic Distance

We have presented in a previous work [26], not published yet, two semantic similarity measures δ and Δ. We have proven their accuracy to establish semantic relatedness and query relevance. We have defined the δ and Δ as: $\delta(A,B) = \frac{\cap_{A+B}}{\cup_{A+B}}$, and $\Delta(D_i,D_j) = \frac{\delta(D_i,D_j)+Jaccard(D_i,D_j)}{2}$, where $\cap_{A+B}$ denotes the sum of the number of occurrences for all the common words in two given texts A and B, $\cup_{A+B}$ denotes the sum of the number of words in A and the number of words in B including eventually their occurrences, and $Jaccard(D_i,D_j)$ denotes the Jaccard similarity measure for two documents D_i and D_j. Now we are interest in a distance function that can measure the relevance and relatedness. The choice of the distance is dictated by the graph representation of the concepts. Let consider the following data represented by Table 1.

Table 1. The Semantic Relatedness between the Documents $D_1, \ldots, D_6$.

Δ	D_1	D_2	D_3	D_4	D_5	D_6
D_1	1					
D_2	0.103	1				
D_3	0.096	0.131	1			
D_4	0	0.291	0.347	1		
D_5	0	0	0.181	0.250	1	
D_6	0	0	0	0.170	0.060	1

We define a distance denoted by d_Δ for all documents $D_i$and D_j such that:

$$d_\Delta(D_i,D_j) = \frac{1-\Delta(D_i,D_j)}{\Delta(D_i,D_j)} \quad (7)$$

Table2. The semantic distances between the documents $D_1, \ldots, D_6$.

d_Δ	D_1	D_2	D_3	D_4	D_5	D_6
D_1	0					
D_2	8.70	0				
D_3	9.41	6.63	0			
D_4	∞	2.43	1.88	0		
D_5	∞	∞	4.52	3	0	
D_6	∞	∞	∞	4.88	15.7	0

We can calculate the distance for all documents D_i, and D_j such that $\Delta(D_i, D_j) \neq 0$ as represented by Table 2.

d_Δ is always positive. When two documents are same the distance is zero. When two documents have no similarity the distance is not defined. d_Δ is a distance but it is not a metric because the triangle inequality is not verified. If the triangle inequality is respected, it could be very important when we have to calculate the path. From now on we only use d_Δ when computing either query to concept relevance or concept to document relatedness.

4. Concept Representation

To represent the concept we have retrieved Wikipedia articles and selected those are complete and well written. The selection is certainly subjective but the selected articles (almost 2.5 millions articles) cover a large range of knowledge if we keep in memory that the current number of English words is represented by 616.500 entrees according to the Oxford English Dictionary, 2nd edition. From each selected article we remove the stop words, apply the stemming and store the remaining in a repository. From each selected article we have only one concept. Concepts are only stored but not indexed. We thus calculate their semantic relatedness with the d_Δ distance and represent them as an undirected graph. The edges are represented by the semantic distances between the articles. If we consider the documents $D_1, \ldots, D_6$ as concepts, we can represent them by an undirected graph as illustrated by Figure 1. When two concepts have no semantic similarity, there is no path from one to the other. By that method we have built an undirected graph of concepts from the selected articles. We can remark that each time we compare two articles the distance is between zero and 500 as long as the stop words are not removed. For this reason we have to remove the stop words and take 500 as the limit to establish the semantic relatedness. The number 500 corresponds to one occurrence of exactly one common word for two documents that have 1000 words as sum of their lengths. Indeed if two entities have a total of one thousand words but less than one word occurs one time in both, we can conclude that they are not semantically related. Consequently, from now on, each time the distance is not less than 500 we conclude that there are neither relatedness nor relevance. We do not need to calculate the relatedness beyond this limit. We thus gain a performance because the computation cost decreases.

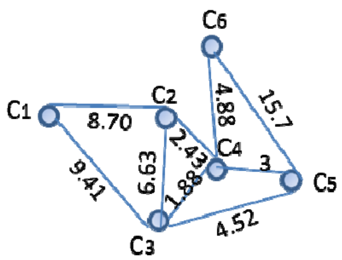

Figure 1. An Undirected Graph of Concepts Constructed from $D_1, \ldots, D_6$.

5. Indexing Documents

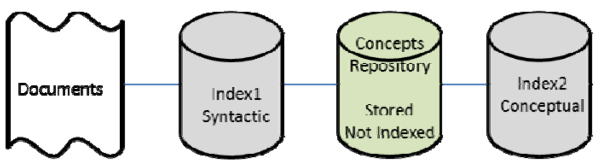

Figure 2. Indexing Representation

The indexing uses Apache Lucene. Thirty three stop words are removed from each document and stemming is applied using the porter algorithm. In addition we have changed the tf-idf similarity measure that uses Lucene. The similarity measure to index the documents is the d_Δ. Lucene is compatible with multi-index. It can easily create and manage multi-index, Fig2. The first way is to index the documents directly using the same d_Δ measure. This index works exactly like syntactic search. Consequently if a document is not related to any concept, it can be retrieved. Our approach extends syntactic search. We thus have two indexes to consider.

The second way to index a document is to measure its relatedness to each of the concepts. Once we have established the semantic relatedness for all the concepts and built the undirected graph of concepts.

Table 3. Indexing Documents

d_Δ	C_1	C_2	C_3	C_4	C_5	C_6
C_1	0					
C_2	8.70	0				
C_3	9.41	6.63	0			
C_4	∞	2.43	1.88	0		
C_5	∞	∞	4.52	3	0	
C_6	∞	∞	∞	4.88	15.7	0
D_1	2	∞	7	∞	∞	∞
D_2	∞	3	∞	∞	∞	9
D_3	∞	∞	∞	4	8	∞

We thus can index any document to be retrieved. If the semantic distance from a document to a concept is less than 500, we add the document to the concept as related document with the corresponding distance. The document is consequently added to the graph of concepts. If we have, for example, three documents D_1, D_2, D_3 and the previous concepts, (section 4) as represented by Table 3, we can index the concepts and add the documents to the graph as represented, Figure 3.

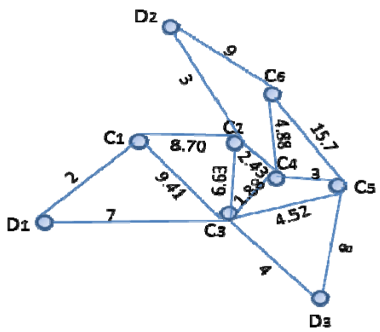

Figure 3. Conceptual Indexing Representation

6. Query Matching

Each query is processed in two different directions according to the index we consider. To search in the index created directly, we process the query as in syntactic search. There is nothing to change and we only call Lucene's Index Searcher to process the query. To search the index that have been built with the concepts, we have to consider the query as a document and measure its relatedness to the concepts. When we know the relatedness of the query to the concepts, we can calculate the distance from the query to the documents via the matched concepts. We thus consider the paths from the query to the documents. If the path to a document is less than 500, the document is returned with the corresponding distance. Otherwise null is returned. Index1 is processed first and returned documents are collected and sent to a renderer. Index2 is processed at the second time, and each retrieved document is checked in the list of retrieved documents from index1. When a document that has been already returned from index1 with a given distance d_1 is again returned from index2 with another distance d_2, we compare the two distances and return the document with the minimum distance $\min\{d_1, d_2\}$ to avoid the no risk of duplication. If a document has not yet been returned from index1, we return the document with its corresponding distance. If we consider the following graph, fig4, where D_i are documents, C_i concepts and Q a query, we can calculate the paths from Q to each of the documents. We thus can retrieve from index2 related documents. Related documents are those within a distance less than 500 from the query Q via their related concepts. To retrieve each relevant document we have to sum the distance from the query to its related concept and the distance from that document to the concept, as indicated by the arrows, Figure 4.

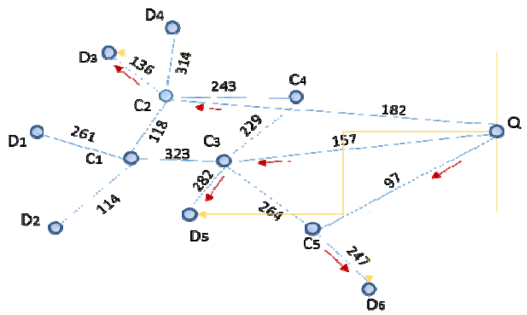

Figure 4. Index2 Query Processing

7. Discussion

The aims of this study, at this step, is to show that one can retrieve documents related to a given query without knowing the magic word that link you to the information needed. The approach extends syntactic search. The first contribution of this methods is to use the same measure to compute both query to concepts and concepts to documents relevance. This oneness allows us to express the path and retrieve the relevant documents. The second contribution is that the results are absolute, not corpus dependent, unlike the works mentioned earlier. The last contribution is to consider the concepts like they are: semantically dependent. The question we expected to answer is to score the improvement providing the rate for both recall and precision. The limitation is that at this step we have not been able to use the concept's relatedness. For example document D_2 (Figure 4) is relatively closed to query Q but, at this step of the implementation, we are unable to retrieve documents that are not directly linked to the concepts matched by the query. For this limitation we did not investigate to measure the accuracy of this method compared to syntactic search. It seems for us more important to develop a method that can retrieve all the relevant documents. In addition, one may ask why the graph representation of the concepts if we do not use that information. At this step the semantic relatedness of concepts have not been used. These issues lead us to investigate query expansion. Query expansion is one the solutions of the interrogations we may have at this step of the implementation.

8. Conclusion

We have presented a concept based approach for information retrieval. Our approach is based on Wikipedia articles. It extends syntactic search using semantic relatedness. It presents another way to improve syntactic search. All the presented concepts are different, and each one is related to only one subject therefore our method overcomes both polysemy and synonymy problems. The semantic measure applied to the graph structure presents an opportunity to better optimize the semantic relevance. Nevertheless the concept's semantic relationships have not yet been in use. Our future work is to increase the performance with the concept-to-concept interactions.

References

[1] Julio Gonzalo, Felisa Verdejo, Irina Chugur, Juan Cigarran Indexing with WordNet synsets can improve text retrieval. *Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP, Montreal.* 1998.

[2] Rada Mihalcea, Dan Moldovan. Semantic Indexing using WordNet Senses. *Semantic indexing using WordNet senses. Proceedings of the ACL 2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval.* 2000.
[3] Anjo Anjewierden, Suzanne Kabel. Automatic indexing of documents with ontologies. *13th Belgian/Dutch Conference on Artificial Intelligence.* 2001.
[4] Jacob Kohler, Stephan Philippi, v Michael Specht, Alexander Rueeg. Ontology based text indexing and querying for the semantic web. *Knowledge-Based Systems.* 2006; 19(8): 744–754.
[5] Rifat Ozcan, Y Aalp Aslandogan. Concept Based Information Access Using Ontologies and Latent Semantic Analysis *Information Technology: Coding and Computing.* ITCC. 2005; 1: 794 – 799.
[6] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*.v 1990; 41(6): 391–407.
[7] Susan T Dumais. Latent Semantic Analysis. *Annual Review of Information Science and Technology.* 2005; 38: 188
[8] Lawrence W Wright, Holly K Grossetta Nardini, Alan R Aronson, Thomas C. Rindflesch Hierarchical concept indexing of full-text documents in the Unified Medical Language System information sources map..*Journal of the American Society for Information Science.* 1999; 50(6): 514-523.
[9] Poshyvanyk D, Marcus A. Combining Formal Concept Analysis with Information Retrieval for Concept *Location in Source Code. Program Comprehension. ICPC '07. 2007: 37-4*
[10] Yang Y, Chute CG. *Words or concepts: the features of indexing units and their optimal use in information retrieval.* Proc Annu Symp Comput Appl Med Care. 1993; 685-9.
[11] Hersh W, R Hickam DH, Leone TJ Words. Concepts or both: optimal indexing units for automated information retrieval. *Annual proceeding of computer applied medical care.* 1992; 644-648.
[12] Concept indexing. Angi Voss, Keiichi Nakata, Marcus Juhnke. *Proceedings of the international ACM SIGGROUP conference on Supporting group work.* 1999; 1-10.
[13] Conceptual Indexing: A Better Way to Organize Knowledge. *Technical Report*. *Sun Microsystems, Inc. Mountain View, CA, USA.* 1997.
[14] Wright, Holly K Grossetta, Nardini, Alar R Aronson, Thomas C. Rindflesch. Hierarchical Concept Indexing of Full-Text Documents in the Unified Medical Language System. Information Sources Map Lawrence W. *Journal of the American Society for Information Science (JASIS*). 1999; 50(6).
[15] Yifan, Gui-Rong Xue, Yong Yu. Advertising keyword suggestion based on concept hierarchy. *Proceedings of the International Conference on Web Search and Data Mining.* 2008; 251-260.
[16] Hersh WR, Greenes RA. SAPHIRE an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Computers and Biomedical Research.* Academic Press Professional, Inc. San Diego, CA, USA. 1990; 23(5): 410-425.
[17] Concept search: Semantics enabled syntactic search. F. Giunchiglia, U. Kharkevich, and I. Zaihrayeu. Proceeding of the 6th European Semantic Web Conf. ESWC. 2009: 429-444.
[18] Placing search in context: the concept revisited. Proceedings of the 10th international conference on World Wide Web. ACM New York, NY, USA. 2001; 406-414.
[19] Philipp Cimiano, Antje Schulz, Sergej Sizov, Philipp Sorg, Steffen Staab*. Explicit vs. Latent Concept Models for Cross-Language Information Retrieval.* Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Pasadena, CA.* 2009: 1513-1518.
[20] Michael Bendersky, W Bruce Croft. *Discovering key concepts in verbose queries*. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM New York, NY, USA. 2008: 491-498
[21] George Karypis, Eui-Hong (Sam) Han. *Fast supervised dimensionality reduction algorithm with applications to document categorization & retrieval*. Proceedings of the ninth international conference on Information and knowledge management. ACM New York, NY, USA. 2000: 12-19
[22] Jasminka Dobša, Bojana Dalbelo Bašic. Comparison of information retrieval techniques: latent semantic indexing and concept indexing. *Journal of Information and Organizational Sciences.* 2004; 28.
[23] Damian Borth, Adrian Ulges, Thomas Michael Breuel. Automatic concept to query mapping for web based concept detector training. Proceedings of the 19th ACM international conference on Multimedia. ACM New York, NY, USA. 2011: 1453-1456.
[24] Caporaso J Gregory, William A Baumgartner Jr, Hyun-min Kim, Zhiyong Lu, Helen L Johnson, Olga Medvedeva, Anna Lindemann, Lynne M Fox, Elizabeth K White, K Bretonnel Cohen, Lawrence Hunter*. Concept Recognition, Information Retrieval and Machine Learning in Genomics Question-Answering*. TREC 2006 Proceedings. 2006.
[25] Hele-Mai Haav, Tanel-Lauri Lubi. *A survey of concept-based information retrieval tools* on the web5th East-European Conference, ADBIS 2001, Vilnius, Lithuania. 2001.
[26] Valuing Semantic Similarity. Abdoulahi Boubacar, Niu Zhendong.