Improving Web Search Results with Explanation-aware Snippets
An Experimental Study

Andias Wira-Alam and Matthäus Zloch
GESIS – Leibniz-Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany

Keywords: Knowledge Extraction, Wikipedia, Information Filtering and Retrieval, Experimentation.

Abstract: In this paper, we focus on a typical task on a web search, in which users want to discover the coherency between two concepts on the Web. In our point of view, this task can be seen as a retrieval process: starting with some source information, the goal is to find target information by following hyperlinks. Given two concepts, e.g. chemistry and gunpowder, are search engines able to *find the coherency and explain it*? In this paper, we introduce a novel way of linking two concepts by following paths of hyperlinks and collecting short text snippets. We implemented a proof-of-concept prototype, which extracts paths and snippets from Wikipedia articles. Our goal is to provide the user with an overview about the coherency, enriching the connection with a short but meaningful description. In our experimental study, we compare the results of our approach with the capability of web search engines. The results show that 72% of the participants find ours better than these of web search engines.

1 INTRODUCTION

The Web is a great resource for everyone. A search process on the Web means finding useful information in an efficient and simple way. It is for us a fascinating problem to investigate explaining *why* two concepts are related to each other and *how* these two concepts can be connected.

The Web evolved to (and still is) a huge knowledge base, where everyone who is connected to it may search for information. But as this huge pot of data grows new methods and algorithms have to be invented to face this vast and increasing availability of data. From the very beginning of the Web, search engines build an interface to the available data. Probably none of the Web search engines nowadays work like the ones from the beginnings, whose first attempts to display adequate results to a users query were finding exact matches in text fragments of one search engine's Website index.

Today, search engines stop at nothing to find more sophisticated methods, e.g. Google Knowledge Graph[1], to display more adequate results. Moreover, a Web search engine is even more integrated in our everyday life, to answer such complex questions, e.g. "Who is the 44th President of the USA?". Beside that, people investigate to find a connection between two terms/concepts that potentially have something in common. A typical scenario is a user reading a newspaper article about some specific topic. At first sight, it is not uncommon that in comprehensive articles or scientific papers readers cannot comprehend the connection between two concepts, mentioned in a sentence or paragraph, especially for young readers.

On the Web, for instance, an article about fireworks might mention chemistry in one sentence and gunpowder in an other, where the direct coherency between the two concepts is not obvious. One might wonder, what does chemistry and gunpowder have in common? As from the information retrieval point of view this challenge can be seen as a retrieval problem, in which users want to discover the coherency between two concepts on the Web.

Based on this assumption, we explore a novel way to build a bridge between the information gaps. Our intention is to provide the users with a quick overview showing the connection of two concepts. By the same analogy with web surfers' behavior, our approach lies on the hyperlinks and the augmenting texts surrounding the hyperlinks. Generally, a web document is structured into logical parts. A web surfer follows the hyperlinks in order to get more detailed explanation about what the surfer is looking for. As an illustration, Figure 1 shows a typical walk of a surfer discover-

[1] http://www.google.com/insidesearch/features/search/knowledge.html

Wira-Alam A. and Zloch M..
Improving Web Search Results with Explanation-aware Snippets - An Experimental Study.
DOI: 10.5220/0004372504590464
In *Proceedings of the 9th International Conference on Web Information Systems and Technologies* (WEBIST-2013), pages 459-464
ISBN: 978-989-8565-54-9

ing the connection between Artificial Intelligence and Semantic Web (in Wikipedia). The text excerpts surrounding the hyperlinks might give a quick overview about the connection without reading through all articles. These texts also indicate which hyperlink will be followed and whether the target information will be found in the end. Our contributions in this paper are as follows: (a)Proposing a novel but efficient way of enriching the search engine results with explanation-aware snippets; (b)Performing user studies to validate our proposed approach. The results show that 72% of the participants find ours better than these of web search engines.

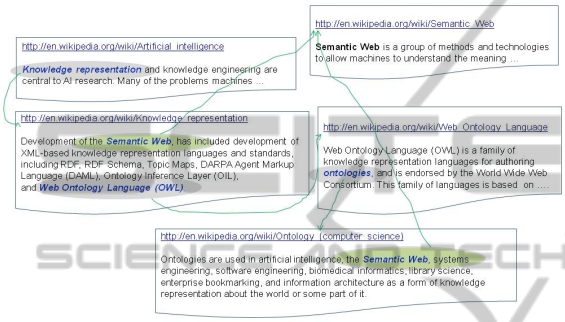

Figure 1: The snippets collection as an expected result in explaining the connection between two concepts: Artificial Intelligence and Semantic Web(Wira-Alam et al., 2010).

2 PROPOSED APPROACH

As mentioned above, the Web is a great learning resource for everyone, especially Wikipedia. The majority of Web users use Wikipedia as their entry point for learning. For instance, if a user is interested in a particular topic, the user may also be interested in its connection to other topics. In order to localize the problem scenarios easier, we focus in this paper on school-related subjects[2] since the participants in the user studies are mainly school students, college students, or graduates. This also helps us to have relevant judgments from the participants.

According to the study presented in (Weller et al., 2010), 95.2% of students from across disciplines use Wikipedia as an entry point to seek information. In our own study, 92.5% of 200 participants stated that they use / have used Wikipedia for learning. Most of the participants in this study have college degrees, 3 have high school diplomas and 17 have attended college with no degree. Based on this fact, we use the English Wikipedia articles as our primary knowledge base.

[2]One can also say e-Learning scenarios.

For Web search engines, a result list is a set of documents. Each result consists of a link and a snippet of a document. In Figure 2, we show the results list provided by Google answering the query "Vitamin C Health". A document may not cover all information needed by users. Some important points may be missing or spread arbitrarily across the found documents. In this matter, the results seem uncomprehensive and therefore it is quite likely that users are not satisfied before clicking of the documents that might answer the query. What if the answers to this query reside in more than one document? Users need to click on each document and read through all the contents in the selected documents.

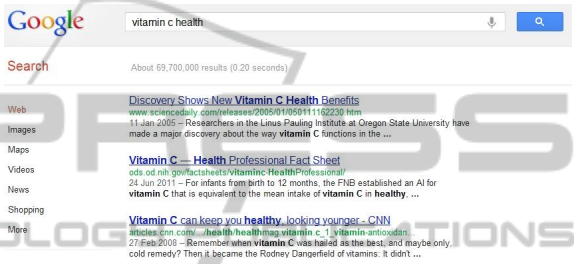

Figure 2: Google's search results for the keywords "Vitamin C Health".

We conduct an initial user study to legitimate this problem before we move to the details of the proposed approach. In addition to the list of documents as provided by Web search engines, we measure the quality of the result provided by our approach by asking 100 participants to rate the following text snippet:

"The richest natural sources are fruits and vegetables, and of those, the Kakadu plum and the camu camu fruit contain the highest concentration of the vitamin. It is also present in some cuts of meat, especially liver. Vitamin C is the most widely taken nutritional supplement and is available in a variety of forms, including tablets, drink mixes, crystals in capsules or naked crystals.

Nutrition (also called nourishment or aliment) is the provision, to cells and organisms, of the materials necessary (in the form of food) to support life. Many common health problems can be prevented or alleviated with a healthy diet."

The results show that 18 participants gave it the best rating and 44 participants the second best. Overall 18 participants rated with "very high" confidence, 57 with "high" confidence, and only 2 rated with "very low" confidence. Thus, we prove our claim that the users' information needs may also be filled in this way.

In order to focus on the user studies to validate our approach, we choose altogether five concept pairs that are considered to be relatively known subjects. The pairs can be seen in Table 1.

Table 1: Concept pairs, paths length, and number of extracted paths.

Concepts	*#Path / Length*
Vitamin C, Health	1 / 2
Mathematics, Computer Science	14 / 2
Chemistry, Gunpowder	8 / 3
Biochemisty, DNA	13 / 2
Computer Science, Bioinformatics	5 / 2

2.1 Extracting Paths and Snippets

We extract all possible paths between concepts / terms based on the hyperlinks-graph, as described in (Wira-Alam and Mathiak, 2012). However, in order to filter out only the relevant articles, we use the selected titles provided by the 2008/9 Wikipedia Selection for schools[3]. Since we only need relevant titles, we ignore articles that are not listed in the selection. This reduces the number of extracted paths drastically[4].

In order to choose the best paths, we first calculate each similarity between two terms using a cosine similarity as follows

$$sim(A,B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

where term vectors A and B are calculated with *tf-idf*. The articles are preprocessed by stripping punctuations and symbols, as well as removing stopwords.

Afterwards, we define a reachability score for each path as follows

$$r_{ij} = \prod_{i}^{j-1} sim(term_i, term_{i+1}) \quad (2)$$

where j is the number of terms in a path. This score describes a "probability" of reaching a target document given a source document.

A snippet is a text excerpt surrounding a hyperlink, which is more than anchor text, using co-occurrence term windows. A window size is in this case a paragraph. The cosine similarity measure tells us which terms are most highly relevant and therefore can be used to score words. Based on the cumulative score of each snippets, we rank the extracted snippets to be shown to the users. The snippets are extracted by using RelWik Extractor[5] (Mathiak et al., 2012).

[3]http://schools-wikipedia.org/

[4]The problem of finding paths is that the time complexity is very high. The worst-case complexity tends to be $O(\bar{n}^{max\_path\_length})$, such that $\bar{n}$ is the average number of outlinks of an article. Currently, it works reasonably with maximum path length of 3.

[5]The RelWik Extractor tool is accessible: http://multiweb.gesis.org/RelationShipExtractor2/

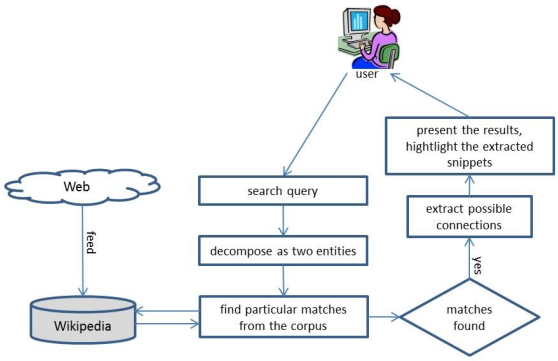

Figure 3: A high-level overview of the proposed approach.

Overall, an overview of the proposed approach is depicted in Figure 3.

3 EVALUATION

First, for the next four pairs, we evaluate the best extracted paths based on the *reachability score* against the popular votes from the participants. For each pair, we asked 100 participants to vote the best path and the best extracted snippets according to their subjective point of view. In Table 2, we see the details of the evaluation for the pair "Mathematics" and "Computer Science". The fourth column of the table shows the ratings for the extracted snippets. Moreover, we also show the agreement between the algorithm and the popular votes from the participants by using Kendall's W score[6]. The other evaluation details can also be found in Table 4, 5, and 6 in the Appendix.

For almost all pairs there is also an interesting fact that some of the snippets extracted from the unfavored paths are rated high. The reasons for this are not obvious and therefore are left for discussion. One of our hypotheses is that unfavored paths share snippets with the favored ones. Another hypothesis, in addition to the background knowledge of the participants, is that the texts could influence the participants' opinion.

Based on the results of the previous experiment, we compare our approach with the results provided by popular search engines[7]. Initially, we start with examining search engine results by giving the 5 concepts we use for the experiment as search queries. We found that to almost all given keywords the Wikipedia articles appeared in the first top 10 results. Only Bing did not show Wikipedia articles in their top 10 list for the keyword "Vitamin C Health".

[6]We ranked the scores of columns 2, 3, and 4, to calculate the Kendall'S scores using this tool: http://www.stattools.net/ KendallW_Pgm.php.

[7]Google, Yahoo, and Bing.

Table 2: Results for the first five paths voted by the users for the pair {Mathematics, Computer Science}. Of 100 participants, 3 are high school students, 11 attend college with no degree, 44 are students in computer sciences, and 10 in mathematics. (Kendall's W score = 0.73).

Paths	*r Score / Votes (%) / Rating*
Mathematics, Applied Mathematics, Computer Science	0.54 / 35 / 3.56
Mathematics, Computer Science	0.69 / 15 / 3.16
Mathematics, Calculus, Computer Science	0.44 / 7 / 3.35
Mathematics, Linear Algebra, Computer Science	0.47 / 7 / 3.39
Mathematics, Operations Research, Computer Science	0.34 / 7 / 0

We perform a second user study to compare the results of search engines and our approach given the pairs as in Table 1. For each pair, we ask 20 participants with related knowledge background to do the following instructions: (a)Given two concepts, as given in Table 1, each participant is asked to find the possible best answer on the Web using the participant's favorite search engine by formulating an own query; (b)Each participant is given time to find the possible best answer on the Web, and then asked to compare the snippets we provided: Which one is better? (c)Finally, the participants are asked to give ratings (from 0 to 10) on how helpful the snippets are as an entry point to understand the relationships between the given terms.

Overall, about 90% of the participants used Google as search engine and point to Wikipedia, research papers / journals, eLearning resources / Wikis, or lecture notes as their destination. Some of the participants that favor our solution also stated that the snippets help to understand and give proper information about the relationship, instead of reading several documents. Otherwise, participants that did not favor our solution stated that the snippets are not well written, difficult to understand, and lack of details. As seen in Table 3, we present an overview about the results. For the given scenarios, 72% of the participants find our results are better than these of web search engines. Besides, our results are also rated as helpful at about 7.5 out of 10, which can be seen as a significant improvement. The whole experiments in this paper had been conducted online. We use Amazon Mechanical's Turk as online platform to recruite the participants for the experiments.

Table 3: Evaluation of comparing our approach with search engines.

Concept Pair	*Better?*	*Helpful?* *(scale: 0-10)*
Mathematics, Computer Science	Yes: 17, No: 3	Median: 8, Mean: 7.9
Chemistry, Gunpowder	Yes: 13, No: 7	Median: 8, Mean: 7.2
Biochemistry, DNA	Yes: 18, No: 2	Median: 7, Mean: 7.45
Computer Science, Bioinformatics	Yes: 10, No: 10	Median: 7, Mean: 7.05

4 RELATED WORK

Previous research explored a variety of different methods to compute semantic relatedness between terms. There is a large number of semantic distances that had been investigated e.g. in (Strube and Ponzetto, 2006; Cilibrasi and Vitanyi, 2007), just to name a few. In (Islam and Inkpen, 2008), the Semantic Text Similarity (STS) has been developed as a variety of the Longest Common Subsequence (LCS) algorithm and a combination of other methods. It is optimized on very short texts, such as single sentences and phrases; it was evaluated by using definitions from a dictionary. Snippets extraction was also studied here (Li et al., 2008); however the authors focused on the variable length of snippets resulting out of a query, not on the relationships.

In (Auer and Lehmann, 2007), the authors make use of the structured information contained in Wikipedia template instances. These templates are analyzed and converted into triples: the Wikipedia page title corresponds to the subject, the template attribute constitutes the predicate and the corresponding attribute value is the object. In addition, class membership of each Wikipedia page is determined. This information can be represented visually as an information map that allows browsing the extracted relations. Alternatively, extracted information can be queried using a graph pattern builder. By querying which relations have been found between two terms, the semantic relatedness between concepts can be determined indirectly. However, since only structured information contained in pattern instances is used, no snippets for explaining the relationships between concepts can be extracted using this approach. Furthermore, only a small part of information concerning two concepts is represented by template instances. Uncommon relations may therefore hardly be found using this method.

The link-based measures have also been applied to graph structures derived from Wikipedia. In (Strube and Ponzetto, 2006), the authors use the Wikipedia category system as the underlying semantic network. (Yeh et al., 2009) uses the links between Wikipedia articles. Articles serve as vertices, links as edges in their graph. Semantic relatedness is then computed by performing random walks with personalized PageRank. Similarly, (Islam and Inkpen, 2008) determines the shortest path on the link structure between articles for semantic relatedness estimation. These methods yield quantitative measures of relatedness but give no insights on how concepts are related to each other qualitatively. Beyond these distance measures, other authors additionally use the anchor texts of links as a knowledge source(Milne and Witten, 2008).

Recently, (Shahaf and Guestrin, 2010; Shahaf et al., 2012b; Shahaf et al., 2012a) introduce a methodology to find a chain of documents that is best suited to guide users from one document to another that describes a related and thematically dependent topic. However, the users need to read through all extracted documents in order to figure out the whole topics. In (Nuzzolese et al., 2011), the authors explain how to use Wikipedia paths popularity in order to describe things or objects. However, they only investigated the paths of length 1. As in (Mathiak et al., 2012), we work in the same direction, however the authors focused only on the concepts that are directly connected. The use of text excerpts have been studied with human judgement about the relationship between terms. In this paper, however, we explore further about the usefulness of text excerpts to solve a retrieval problem, in particular for the concepts that are not necessarily directly connected.

5 CONCLUSIONS AND FUTURE WORK

We see that explaining the relationship between concepts in an automatic way, by displaying clear and logic paths and text snippets as an overview, is a novel problem. In this paper we describe our approach shortly and examine this method that we have implemented as a prototype-system. The experiments we have made show examples of extracted paths together with snippets and user ratings, which were conducted in several user studies. Furthermore, comparing the user ratings with our implemented scoring function, we were able to shed some light on the quality of it.

We let the users compare our results to those obtained by their favored web search engines and gained positive feedback. It points out that, in the majority of the cases the rating for paths and snippets was not equal to those computed by our scoring function and these of the users. But obviously, the majority of the users find the extraction of the paths and text snippets helpful, as shown in Table 3. However, based on the current results, we have to investigate further to develop a better ranking algorithm.

The experiments arranged in this paper lead us to the following conclusions: web users do find this method of connecting two unknown concepts by paths and short text snippets useful. We received good feedback for the extraction of some example concepts. Since there are only a few researchers addressing this problem, we believe that our work can contribute to the development of the World Wide Web, particularly the Information Retrieval model on Web search.

However, our approach has also limitations, since we currently only consider terms found in Wikipedia and there is no suggestion for term disambiguation. So far, our algorithm works with Wikipedia URL's and topics. It would be promising to investigate in a more generic API, thus it could be applied to other document corpora. We believe that our approach would be suitable for more general scenarios in the Web search, thus we plan to build an interactive user interface for leveraging user feedback to refine the results. Moreover, we also plan to further investigate our approach with other domain-specific document corpora for a more extensive user study.

ACKNOWLEDGEMENTS

We thank Víctor Manuel Martínez Peña and Sigit Nugraha for the implementation and enhancement of the tool RelWik Extractor. We also thank Matthias Krautz and colleagues for the fruitful discussions as well as the useful inputs and advice on improving this work.

REFERENCES

Auer, S. and Lehmann, J. (2007). What have Innsbruck and Leipzig in common? extracting semantics from wiki content. In *Proceedings of the 4th European conference on The Semantic Web: Research and Applications*, ESWC '07, pages 503–517.

Cilibrasi, R. L. and Vitanyi, P. M. B. (2007). The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383.

Islam, A. and Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data*, 2(2):10:1–10:25.

Li, Q., Candan, K. S., and Yan, Q. (2008). Extracting relevant snippets for web navigation. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, AAAI'08, pages 1195–1200.

Mathiak, B., Martínez Peña, V. M., and Wira-Alam, A. (2012). What is the relationship about? - extracting information about relationships from wikipedia. In *WEBIST*, pages 625–632.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08.

Nuzzolese, A. G., Gangemi, A., Presutti, V., and Ciancarini, P. (2011). Encyclopedic knowledge patterns from wikipedia links. In *Proceedings of the 10th international conference on The semantic web - Volume Part I*, ISWC'11, pages 520–536.

Shahaf, D. and Guestrin, C. (2010). Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 623–632.

Shahaf, D., Guestrin, C., and Horvitz, E. (2012a). Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 1122–1130.

Shahaf, D., Guestrin, C., and Horvitz, E. (2012b). Trains of thought: generating information maps. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 899–908.

Strube, M. and Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, AAAI'06, pages 1419–1424.

Weller, K., Dornstäer, R., Freimanis, R., Klein, R., and Perez, R. (2010). Social software in academia: Three studies on users. acceptance of web 2.0 services. In *WebScience 2010, Raleigh, USA, 2010.*

Wira-Alam, A. and Mathiak, B. (2012). Mining wikipedia's snippets graph: First step to build a new knowledge base. In *First International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data, Heraklion, Greece, 2012.*

Wira-Alam, A., Zapilko, B., and Mayr, P. (2010). An experimental approach for collecting snippets describing the relations between wikipedia articles. In *WebScience 2010, Raleigh, USA, 2010.*

Yeh, E., Ramage, D., Manning, C. D., Agirre, E., and Soroa, A. (2009). Wikiwalk: random walks on wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs-4, pages 41–49.

APPENDIX

In this Appendix, we show the evaluation details for term pairs: "Chemistry" and "Gunpowder", "Biochemistry" and "DNA", and "Computer Science" and "Bioinformatics".

Table 4: Results for the first five paths voted by the users for the pair {Chemistry, Gunpowder}. Of 100 participants, 5 are high school students, 13 attend college with no degree, 10 are students in chemistry and 6 in physics. (Kendall's W score = 0.762).

Paths	*r Score / Votes (%) / Rating*
Chemistry, Chemical Reaction, Potassium Nitrate, Gunpowder	0.22 / 38 / 3.45
Chemistry, Chemical Reaction, Sulfur, Gunpowder	0.23 / 19 / 3.37
Chemistry, Chemical Bond, Nitrogen, Gunpowder	0.21 / 9 / 3.44
Chemistry, Oxygen, Nitrogen, Gunpowder	0.29 / 5 / 0
Chemistry, Sodium Chloride, Iodine, Gunpowder	0.21 / 5 / 0

Table 5: Results for the first five paths voted by the users for the pair {Biochemistry, DNA}. Of 100 participants, 7 are high school students, 11 attend college with no degree, 7 are students in chemistry, 11 in biology, and 3 in medicine. (Kendall's W score = 0.49).

Paths	*r Score / Votes (%) / Rating*
Biochemistry, Genetic_code, DNA	0.58 / 44 / 3.19
Biochemistry, Cell (biology), DNA	0.51 / 25 / 3.59
Biochemistry, DNA	0.74 / 12 / 2.86
Biochemistry, Organism, DNA	0.46 / 10 / 0
Biochemistry, Antibody, DNA	0.35 / 4 / 0

Table 6: Results for the first five paths voted by the users for the pair {Computer Science, Bioinformatics}. Of 100 participants, 4 are high school students, 11 attend college with no degree, 35 are students in computer science and 10 in biology. (Kendall's W score = 0.29).

Paths	*r Score / Votes (%) / Rating*
Computer Science, Computer Programming, Bioinformatics	0.61 / 28 / 2.64
Computer Science, Information, Bioinformatics	0.51 / 24 / 3.05
Computer Science, Computational Chemistry, Bioinformatics	0.44 / 17 / 3.42
Computer Science, Statistics, Bioinformatics	0.45 / 16 / 0
Computer Science, Bioinformatics	0.78 / 15 / 0