Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 9 (2013) 1 – 10

Procedia Technology

CENTERIS 2013 - Conference on ENTERprise Information Systems / PRojMAN 2013 - International Conference on Project MANagement / HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies

A Fuzzy Algorithm for Optimizing Semantic Documental Searches

Sara Paiva*

sara.paiva@estg.ipvc.pt; Viana do Castelo Polytechnic Institute, Portugal

Abstract

Search for documents is a common and pertinent task lots of organizations face every day as well as common Internet users in their daily searches. One specific document search is scientific paper search in reference manager systems such as Mendeley or IEEExplore. Considering the difficult task finding documents can sometimes represent, semantic search is currently being applied to improve this type of search. As the act of deciding if a document is a good result for a given search expression is vague, fuzziness becomes an important aspect when defining search algorithms. In this paper, we present a fuzzy algorithm for improving documental searches optimized for specific scenarios where we want to find a document but don´t remember the exact words used, if plural our singular words were used or if a synonym was used. We also present the application of this algorithm to a real scenario comparing to Mendeley results.

Keywords: Semantic search; Documental search; Fuzzy algorithms.

* Corresponding author. Tel.: +351 258 819 700; fax: +351 258 827 636.
E-mail address: sara.paiva@estg.ipvc.pt

doi:10.1016/j.protcy.2013.12.001

1. Introduction

Search for information continues to be a current matter to address and to improve. Search necessities keep increasing as they apply to several and distinct domains such as finding documents in a company's intranet, scientific articles in specific sites or the most common broad general information search on the World Wide Web. In all these cases, users always want the same thing: quickly find what they are looking for. We have seen in the past traditional IR techniques and their limitation to some type (majority) of search needs. Semantic search appeared to solve some of those limitations by annotating resources with meta-data and including their meaning to generate better results.

We have dedicated some investigation work on this theme, namely when developing the systems PRECISION and GSSP. PRECISION [1], [2] stands for "guided and PeRsonalized Expression ConstructIon with Semantic validation" and is a guided-based search system with two main characteristics: semantic validation and personalized natural language generation of search expressions. Additionally, the system, which is oriented to comparative searches, supports 1:N ontology class relations and also the notion of search and auxiliary classes which gives each of these type of class different roles in the query construction process. GSSP [3] stands for "Generic Semantic Search Platform" and its main goal is to provide a platform that allows a given search system to incorporate semantics in its search process. GSSP is built on top of PRECISION and was designed to suite any scenario where searches are helpful and with few configuration needs. GSSP intends to be easy, not only to the end user, but also for the system that is trying to adopt it to its use. GSSP adds other input methods when comparing with PRECISION - the free input method. This type of input method is probably the most complicated one to satisfy as one of the main difficulties in obtaining a good precision in the search results is to understand clearly what the user wants and then have the algorithms to find that information. When it comes to understanding what the user wants, free search can be very tricky as the same word can have different contexts and plural/singular or different verbs conjugation can make a big difference. So far as GSSP goes, we defined four criteria for handling free searches that include reducing each word of the search expression to its origin and then comparing them to the description of existing resources, to the description of related resources and to resources that use synonyms.

Following this line of investigation, we believe some degree of fuzziness is missing to the already defined criteria as deciding that a given resource satisfies a given search expression is not entirely black and white as it is a vague decision. As a response to that, in this paper we present a fuzzy algorithm for optimizing semantic documental searches. In concrete, we try to address the following scenario:

> *Bob is a master degree student currently writing his dissertation on Education Sciences. For this task, he feels the need to use several scientific papers digital repositories such as Mendeley or IEEExplore. Two days ago, during his searches, he found an interesting paper but it was time to go home and he closed the Internet browser. Late that night, he made a routine cleanup task on his computer erasing internet temporary files. Yesterday, he picked up where he left and remembered that paper he saw and went back to Mendeley to search for the paper but was having some troubles finding it. He remembered the paper had somewhere the words "students" and "learn" in the title but he didn't know for sure if those were the exact words used, or if synonyms were used instead, other verb conjugation ("learning"), etc. It took him several search expressions to finally find the paper, which was ranked too low.*

The rest of this paper is organized as follows: the next section introduces a literature review on search techniques. Section 3, to justify the need and demonstrate the pertinence of the concrete issue we are addressing, presents some analysis on the results obtained when performing searches on a reference manager – we chose Mendeley. In section 4 we present the results of the experiments and evaluation made with the algorithm. Finally, we present the conclusions of this work.

2. Related concepts background

Information Retrieval Systems (IRS) first appeared to store information necessary to the increasing search needs that the computer science development growth made possible. They were first used to manage the literary scientific explosion occurred in the middle XX century [4]. Currently, IRS systems are oriented to obtaining documents that contain relevant information for the user considering the search he made. The definition provided by Manning [5] traduces exactly this: obtaining information and searching for material (usually documents) of non-structured nature (usually text) from a collection (usually stored in computers) that fulfills a given need for information.

IRS concept was created by Calvin Moores around 1950. In the end of that decade, Luhn [6] became the first to apply automatic indexing by concluding that the frequency with which words appear in a document can be used to define a relevance degree. Salton and Yang [7] gave an important contribution in this field when they find out that terms that appear with too much frequency are not the most representative but those that appear with a medium frequency. For that, they used the concept of stopword list. On another hand, Crouch and Yang [8] were the first to develop an automatic thesaurus using keywords that could be used to index documents and perform searches. Regarding this, in the beginning of the 70´s, Bely, Borillo, Virbel and Siot-Decauville [9] made the first automatic thesaurus using document abstracts. In 1971, the first work of information recovery was done by Sparck Jones [10] that used measures of association between keywords based on occurrence frequency. The concept of relevance feedback in IRS, that consists on using in the next iteration the result of the current one, is due to Rocchio [11]. In 1984, Jones e Tait [10] presented search expansion as a way to obtain several other searches that expressed the same need as the original one.

The advent of the Internet potentiated the use of these systems, and IRS started to be applied to large volumes of data. On another hand, this advent showed some limitations of IRS such as the inability to work with unstructured data over which searches become complex what compromises objectivity. In recent years, it was found that even with the best indexing techniques, a good precision in search results cannot yet be obtained. The proliferation of the use of natural language and its processing did not help, as the best performance of these techniques is achieved with reduced dimensions texts. Also, natural language introduces a considerable amount of ambiguity in understanding what the user wants with a given search expression.

In current search systems, as for example Google, it is very easy to get lost or obtain irrelevant information for a given search we made. In reality, it is sometimes too complicated to objectively find specific type of information such as “papers written by Eric Miller”. This search is very specific for a human but is not for a machine as it doesn´t know what an article is or who is Eric Miller. Most likely, we will obtain several documents where the name “Eric” appears or “Miller” (it is what we can hope in case traditional IRS are used) One of the reasons for obtaining this results is the fact that the indexing process is based on the frequency that the terms appear in documents and completely forgetting the notion of context and meaning. That is, however, inevitable once the majority of the contents we find nowadays on the Internet is supposed to be read and interpreted by humans and not to be manipulated by computers.

To improve the search for information it is necessary that search engines can understand what the user wants so they are able to answer objectively. To achieve that, one of the necessary things is that the resources have information that can be helpful to searches such as, for example, an author name, creation data, etc.

The evolution to Web 3.0 – Semantic Web – proposed to clarify the meaning of resources by annotating them with metadata [12] – data over data. By associating metadata to resources, semantic searches can be significantly improved when compared to traditional searches [13]. The last and main goal to reach with semantic search is to allow users the use of natural language to express what he wants to find. Examples of searches are: “what is the age of Madonna?” or “what are the books written by Nicholas Sparks?”.

Several proposals of semantic search systems exists nowadays [14] [15] [16] and studies/investigation on their performance keep being conducted such as in [17] where the authors evaluate the precision of results

obtained with Google, Yahoo, Msn and Hakia. Regarding document search, several proposed can also be found. Nyamsuren and Choi [18] propose the creation of a semantic model of the document, an ontology-like structured semantic annotation of the document with support for structured querying. On another hand, Chatvichienchai and Tanaka [19] refer to the problematic of finding digital documents in an office large repository. The authors present a *technique that collects search terms and their semantic relationship from the documents of some office applications to generate the XML-based search indices that can effectively locate the office documents.* Finally and important to this literature review are fuzzy search engines[20] [21] [22].. The authors seem to agree that synonyms and similar keywords are not taken into consideration in traditional searches, users may need several keywords individually to complete a search or even that all keywords are treated with the same importance.

3. Search results analysis in current document repositories

In order to justify and better explain the issue we are addressing, we started by making an analysis of how document search currently works in a reference manager system: Mendeley [23]. As we previously mentioned, our main goal is to address a specific type of search where we are looking for a specific paper we saw yesterday that had something to do with "students" and "learn" but we don´t know exactly if those were the exact words used, or if synonyms were used instead, other verb conjugation ("learning"), etc.

3.1. Mendeley Reference Manager

For the mentioned purposes, we started by picking a random paper: "Pupil Experiences and Pupil Learning in the Elementary Classroom: An Illustration of a Generative Methodology" [24] and defined six different search expressions using the main keywords of the original paper or variations of those keywords, that could probably be used by someone finding this specific paper that didn´t remember the exact name. Table 1 shows the ranking that the original paper was returned in (results obtained on 17th March 2013) for each search expression used and also the distribution of keywords on the top papers returned.

Table 1. Mendeley results analysis considering variations of a search expression

Search expression	Rank	Docs with 1st term only	Docs with 2nd term only	Docs with both terms	Docs with no term
Pupil	48	33	N/A	N/A	2
Pupil experience	Above 50	29	4	5	11
Pupil experiences	2	1			
Pupil learning	1				
Pupil learn	Above 50	19	10	7	14
Student experiences	Above 50	21	3	22	4

From the analysis of this data we can conclude that 1) a change from **plural to singular** of a given word has a big impact on the returned results as the simple use of the singular version of the second word of the paper made it disappear from the top 50 results to the query; 2) changing the **verb conjugation** (learning to learn) is also enough to exclude the document from the top 50 results; 3) another limitation of the search process is the use of **synonyms** (*students* instead of *pupils*) which shows semantics inexistence; 4) in the majority of the search expressions, some of the results include only a single word of the search expression, or even none. In this last case, after verifying, we conclude that these documents are returned because the words used in the search expressions are present in the abstract. The problem with trusting words in the abstract can

be that they are out of context while relying on the words of the title is more specific and representative of the document´s content.

4. Proposed fuzzy algorithm for documental search

In this section we start by presenting the used methodology and next the proposed algorithm.

4.1 Methodology

Considering the analysis presented in the previous section, we defined as a requirement that the search algorithm is able to handle variations at three main levels: (1) plural versus singular words; (2) difference in verb conjugation; (3) use of synonyms.

As described in [3], the first two variations are dealt with word reduction to its origin using official Language Thesaurus. The issue related with synonyms is solved with keywords expansion also using a Thesaurus so we can maintain a database of keywords/synonyms. At this point, it became fundamental to decide about priorities, which defines the fuzziness component of the algorithm. We believe the existence of the word in the title of the document exactly as it was written in the search query should have the highest priority as it expresses exactly what the user want. This way, we assure that if he knows exactly what he wants, he will obtain it immediately. Next, we consider that eliminating plurals and reducing words to its origin can help in the document retrieval processes. An example would be to return the document "The Expansion of Information Concept" if the user enters the search expression with the keywords "Information Concept**s**". We also eliminate verb conjugations and reduce it to its main form ("walking" is reduced to "walk". An example of this would be to return the document "Effects of information processing on post realistic job preview perceptions" if the user enters the search expression with the keywords "process job". Finally, we consider word synonyms. An example of this would be to return the document "What about my laptop?" if the user enters the search expression with the keywords "computer".

For the process to be possible we have a database with all searchable document titles, which are all submitted to a processing stage: [STEP1] Remove stopwords. ; [STEP2] All remaining words of the title are expanded (synonyms creation) using Altervista (http://thesaurus.altervista.org/) and Priberam (http://www.priberam.pt/); [STEP3] All remaining words of the title are reduced to its origin and then also expanded. With these steps, and for each title, we now have: 1) a collection of words and synonyms of each; 2) a collection of reduced words and synonyms for each.

4.2 The algorithm

Before presenting the criteria, let us admit the following:

1. Let TS be the set of terms of the search expression
2. Let TSR be the set of terms of the search expression, reduced to their origin.
3. Let DC be the set of documents that can be searched
4. Let DC(t) be the document that has the title *t*
5. Let S(DC(t)) be the set of terms that constitute the title of DC(t)
6. Let SR(DC(t)) be the set of reduced terms that constitute the title of DC(t)

Let us also consider the following variables:

7. The number of words of the search expression that are part of a given document with title *t*

$$nts(t) = card\ (TS\ \cap\ S(dc(t)))$$

8. The number of reduced words in the search expression that are part of a given document with title *t*

   $ntsr(t) = card(TSR \cap SR(dc(t)))$
9. The number of synonyms in the set of reduced terms, given by *nsyn(t)*
10. The number of words of the search expression, given by *nwords*

Finally, let us consider the following weights:

11. The weight *wdir* of having words of the search expression directly in the document title
12. The weight *wred* of having reduced words of the search expression in the reduced document title
13. The weight *wsyn* of having synonyms of words of the search expression in the document title

Based on these variables, we calculate the weight of each document (*wdoc(t)*) and documents are then presented to the user ordered by their weight:

*wdoc(t) = (nts(t) * wdir / nwords) + (ntsr(t * wred / nwords) + (nsyn(t) * wsyn / nwords)*

With this algorithm we give more importance to documents that have the biggest number of direct words of the search expression in the document title, followed by the documents that have the biggest number of reduced words and then finally documents that have the biggest number of synonyms. Of course combinations are also possible such as documents that have one direct word and one synonym or document that have 2 reduced words and two synonyms.

Experimental Results

In this section we present some experimental results regarding the application of the algorithm in a real scenario. In concrete, we used the search expressions defined in Table 1 and registered the rank the document "Pupil Experiences and Pupil Learning in the Elementary Classroom: An Illustration of a Generative Methodology" was returned in on Mendeley and with the proposed algorithm. After some experiments, we fixed the weighs in: *wdir = 1*; *wred = 0,85* and *wsyn = 0.7*.

The first search expression tested was "pupil learn" and the results are presented in

Table 2.The weight of the paper with title *t* equal to "Pupil Experiences and Pupil Learning in the Elementary Classroom: An Illustration of a Generative Methodology" was calculated as follows:

*wdoc(t) = (nts(t) * wdir / nwords) + (ntsr(t * wred / nwords) + (nsyn(t) * wsyn / nwords)*
*wdoc(t) = (1*1/2) + (1*0,85/2) + (0*0,7/2) = 0,925*

In this case, the paper we wish to find ranked above 50 with Mendeley algorithm and ranks 2 with our algorithm.

Table 2. Algorithm experimental results with search expression "pupil learn"

Search expression: pupil learn / *nwords* = 2						
Mendeley rank	Document title	*nts(t)*	*ntsr(t)*	*nsyn(t)*	*wdoc(t)*	Algorithm rank
1	The computerized **pupil**.	1			0,5	3
2	National Cultural Values and Their Role in **Learning** : A comparative ethnographic study of state primary schooling in England and		1		0,425	6
3	**Learning** Stations In The Social Studies.		1		0,425	6
4	Understanding metacognition through the use of **pupil** views templates: **Pupil** views of Learning to **Learn**	2			1	1
5	Academic outcomes in school classes with markedly disruptive **pupils**		1		0,425	5
6	Discover: Helping teachers to discover the pleasure of **learning** and teaching		1		0,425	6
7	School finance and opportunities to **learn**: Does money well spent enhance students' achievement?	1			0,5	3
8	Wow factor.	0	0	0	0	10
9	Getting Started with Mendeley NOW !	0	0	0	0	10
10	The quality of **learning**: assessment alternatives for primary education		1		0,425	6
…..	……..	……..	……..	……..	……..	……..
Above 50	**Pupil** Experiences and **Pupil** Learning in the Elementary Classroom: An Illustration of a Generative Methodology	1	1		0,925	2

The second search expression tested was “student experiences” and the results are presented in Table 3. This search shows the combination of direct words, reduced words and synonyms. The paper we wish to find ranked above 50 with Mendeley algorithm and ranks 6 with our algorithm.

Table 3. Algorithm experimental results with search expression "student experiences"

Search expression: student experiences / *nwords* = 2						
Mendeley rank	Document title	*nts(t)*	*ntsr(t)*	*nsyn(t)*	*wdoc(t)*	Algorithm rank
1	Examining The Effects Of **Student** Involvement On African American College **Student** Development	1			0,5	7
2	A Structural Model Of Perceived Academic, Personal, And Vocational Gains Related To College **Student** Responsibility	1			0,5	7
3	The Community College **Student Experiences** Questionnaire: Introduction And Application	2			1	1

4	This Is Who I Am: **Experiences** Of Native American **Students**	1	1		0,925	5
5	The Contribution Of Field **Experiences** To **Student** Primary Teachers' Professional Development	2			1	1
6	University Libraries And **Student** Engagement	1			0,5	7
7	Menehunes In The Library.	0	0	0	0	11
8	Critical Voices In School Reform: **Students** Living Through Change		1		0,425	10
9	Graduate **Student Experiences** At Illinois	2			1	1
10	Developing A Positive Supervision Framework From Negative **Student** Experiences	2			1	1
…..	……..	……..	……..	……..	……..	……..
Above 50	**Pupil Experiences** and **Pupil** Learning in the Elementary Classroom: An Illustration of a Generative Methodology	1		1	0.85	6

Table 4 shows a comparison between Mendeley´s rank and our algorithms when searching for the paper "Pupil Experiences and Pupil Learning in the Elementary Classroom: An Illustration of a Generative Methodology". The results show a significant improvement on the rank the paper was returned in.

Table 4. Comparison between Mendeley´s rank and our algorithms

Search expression	Mendeley rank	Algorithm rank	Observations
Pupil	48	1	Rank 1 along with so many others
Pupil experience	Above 50	3	
Pupil experiences	2	1	
Pupil learning	1	1	
Pupil learn	Above 50	2	
Student experiences	Above 50	6	

Conclusions

The search activity is a common task in most people´s daily routine. Document search is a specific type of search very used in organizations and also by common Internet users. One document search example is scientific paper search using reference manager systems such as Mendeley or IEEExplore. As the search task is a continuously field of research, semantic search has been being applied also to document search as a way to improve results. When defining algorithms, fuzziness is often considered to respond to the vagueness when deciding if a document is a good result for a given query.

In this paper, we presented a fuzzy algorithm for improving documental searches optimized for specific scenarios where we want to find a document but don´t remember the exact words used, if plural or singular words were used or if a synonym was used. The defined algorithm takes into consideration: 1) the

number of direct words of the search expression that are in the document title; 2) the number of word variation (plural/singular or different verbs conjugation) of the search expression that are in the document title; 3) the number of synonyms of the words in the search expression that are in the document title; weights to each one of this components as the fuzziness part of the algorithm.

We started by making an analysis on how easy we could find a specific paper *X* on Mendeley using several search expressions. Then we compared Mendeley´s ranking and our algorithms. From the experimental results the main conclusion is that the algorithm seems to perform well when at least two keywords are used in the search expression. The experiments with one keyword only did not show any improvement on the rank of paper *X*. However, with two keywords, the results are significantly better. With three search expressions that Mendeley´s rank was above 50, our algorithm ranked positions 2, 3 and 6.

As a continuation of this work and to improve the algorithm, we are making experimental tests with other reference manager systems and with larger volumes of data.

References

[1] S. Paiva, M. Ramos-Cabrer, and A. Gil-Solla, "Semantic Query Validation in Guided-Based Systems: assuring the construction of queries that make sense," in Proceedings of the 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2010), 2010, no. 1, pp. 9–14.

[2] S. Paiva, M. Ramos-Cabrer, A. Gil-Solla, A. Fernandez-Vilas, and R. Diaz-Redondo, "Precision : a Guided-Based System for Semantic Validation and Personalized Natural Language Generation of Queries," in Proceedings of the 29th International Conference on Consumer Electronics (ICCE 2011), 2011, pp. 503 – 504.

[3] S. Paiva, M. Ramos-Cabrer, and A. Gil-Solla, "GSSP - A Generic Semantic Search Platform," in 4th Conference of ENTERprise Information Systems – aligning technology, organizations and people (CENTERIS 2012), 2012, pp. 388–396.

[4] M. K. Buckland, "What is a 'document'?," Journal of the American Society of Information Science, vol. 48, no. 9, pp. 804–809, 1997.

[5] C. D. Manning, P. Raghavan, and H. Schutze, An Introduction to Information Retrieval, no. c. 2009.

[6] C. J. Van Rijsbergen, Information Retrieval. 1979.

[7] G. Salton and C. S. Yang, "On The Specification Of Term Values In Automatic Indexing," Journal of Documentation, vol. 29, no. 4, pp. 351–372, 1973.

[8] C. J. Crouch and B. Yang, "Experiments in automatic statistical thesaurus construction," in SIGIR 92 Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, 1992, pp. 77–88.

[9] N. Bely, A. Borillo, J. Virbel, and N. Siot-Decauville, Procédures d'analyse sémantique appliquée à la documentation scientifique. 1970.

[10] K. S. Jones and J. I. Tait, "Automatic Search Term Variant Generation," Journal of Documentation, vol. 40, no. 1, pp. 50–66, 1984.

[11] J. J. Rocchio, "Document retrieval systems - optimization and evaluation - Report ISR-10 to National Science Foundation," Harvard University, 1966.

[12] K. Anyanwu and A. Seth, "The ρ-Operator : Enabling Querying for Semantic Associations on the Semantic Web .," in WWW '03 Proceedings of the 12th International Conference on World Wide Web, 2003, pp. 115–125.

[13] R. Guha, R. McCool, and E. Miller, "Semantic Search," in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 700 – 709.

[14] T. Tran, D. M. Herzig, and G. Ladwig, “SemSearchPro – Using semantics throughout the search process,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 9, no. 4, pp. 349–364, Dec. 2011.
[15] Z. Liu and Y. Zhang, “Research and Design of E-commerce Semantic Search,” 2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 332–334, Nov. 2010.
[16] C. Lv, T. Kobayashi, K. Agusa, K. Wu, and Q. Zhu, “Image Semantic Search Engine,” 2009 First International Workshop on Database Technology and Applications, pp. 156–159, Apr. 2009.
[17] D. Tümer, M. A. Shah, and Y. Bitirim, “An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia,” 2009 Fourth International Conference on Internet Monitoring and Protection, pp. 51–55, 2009.
[18] E. Nyamsuren and H. Choi, “Building a semantic model of a textual document for efficient search and retrieval,” in 11th International Conference onAdvanced Communication Technology, 2009, pp. 298–302.
[19] S. Chatvichienchai and K. Tanaka, “An Effective Document Search Technique by Semantic Relationship Approach,” 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, pp. 53–58, 2009.
[20] L.-F. Lai, C.-C. Wu, P.-Y. Lin, and L.-T. Huang, “Developing a fuzzy search engine based on fuzzy ontology and semantic search,” 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), pp. 2684–2689, Jun. 2011.
[21] R. Li, K. Wen, Z. Lu, X. Sun, and Z. Wang, “An improved semantic search model based on hybrid fuzzy description logic,” 2006 Japan-China Joint Workshop on Frontier of Computer Science and Technology, pp. 139–146, Nov. 2006.
[22] F. P. Romero, J. A. Olivas, J. De Mata, and C. Carmen, “BUDI : Architecture for Fuzzy Search in Documental Repositories,” vol. 16, pp. 71–85, 2009.
[23] Mendeley, “Mendeley Reference Manager.” [Online]. Available: http://www.mendeley.com/.
[24] A. Alton-Lee and G. Nuthall, “Pupil Experiences and Pupil Learning in the Elementary Classroom: An Illustration of a Generative Methodology,” Teaching Teacher Education, vol. 6, no. 1, pp. 27–45, 1990.