International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 6, Issue 3 [March 2017] PP: 28-33

Attribute-Driven Approach for Handling Fuzzy Terms in Domain-Specific Web Search Engines

Dae Young Choi
Department of MIS, YUHAN University, South Korea

Abstract:-In contrast to general-purpose Web search engines, which attempt to index large portions of the WWW using a web crawler, domain-specific Web search engines typically use a focused crawler that attempts to index only Web pages that are relevant to a pre-defined topic or set of topics. Domain-specific Web search engines are becoming increasingly popular because they offer increased accuracy and extra features not possible with general-purpose. However, commercial domain-specific Web search engines still keyword-based and thus generally return too many search results or irrelevant to user's search intentions. It is mainly derived from the inappropriateness on reflecting user's search intentions. To make personalized search results by reflecting user's search intentions appropriately, we propose an attribute-driven approach for handling fuzzy terms in domain-specific Web search engines. Handling fuzzy query in domain-specific Web search engines is one of the most difficult problems. It is mainly derived from the complexity and the degree of freedom of natural language. To reduce the complexity and the degree of freedom of fuzzy query in domain-specific Web search engines, we propose an attribute-driven approach for handling fuzzy terms. It makes personalized search results by reflecting user's search intentions appropriately in domain-specific Web search engines.

Keywords: *Attribute-driven, Domain-specific Web search engines, Personalized search, User's search intentions*

I. INTRODUCTION

Domain-specific Web search engines focus on a specific segment of online content. The domain-specific content area may be based on topicality, media type, or genre of content. Common domain-specific Web sites include shopping, the automotive industry, legal information, medical information, scholarly literature, travel, etc. Domain-specific Web search engines generally focus on one area of knowledge, creating customized search experiences. Commercial examples of domain-specific Web search engines include Shopping.com, Hotels.com, Trulia.com, Pricegrabber.com, Yelp.com, etc. In contrast to general-purpose Web search engines, which attempt to index large portions of the WWW using a web crawler, domain-specific Web search engines typically use a focused crawler that attempts to index only Web pages that are relevant to a pre-defined topic or set of topics. Some domain-specific search sites focus on individual verticals, while other sites include multiple vertical searches within one search engine. Domain-specific Web search offers several potential benefits over general-purpose Web search engines as follows : Greater precision due to limited scope, leverage domain knowledge including taxonomies and ontologies [5], support of specific unique user tasks. In addition, while a general-purpose Web search engines should index all pages, and might use breadth-first search to collect documents, domain-specific Web search engines need only index a small subset. Domain-specific Web search engines are becoming increasingly popular because they offer increased accuracy and extra features not possible with general-purpose Web search engines. In 2013, consumer price comparison websites with integrated vertical search engines such as FindTheBest drew large rounds of venture capital funding, indicating a growth trend for these applications of vertical search technology [9].

Although the domain's limited corpus and clear relationships between concepts provide extremely relevant results for searchers in commercial domain-specific Web search engines such as Pricegrabber.com, Trulia.com, etc, they use still keyword-based search. Thus, they have difficulty in handling fuzzy matching or synonym matching. To enhance the capability of matching in domain-specific Web search engines, a fuzzy search [10] based on a fuzzy matching program need to be considered. Fuzzy search returns a list of results based on likely relevance even though search argument words and spellings may not exactly match. Exact and highly relevant matches appear near the top of the list. Subjective relevance ratings, usually as percentages, may be given. A fuzzy matching program can return hits with content that contains a specified base word along with prefixes and suffixes. For example, if 'planet' is entered as a search word, hits occur for sites containing words such as 'protoplanet' or 'planetary.' The program can also find synonyms and related terms, working like an online thesaurus or encyclopedic cross-reference tool. For example, if the word 'galaxy' is entered, hits are returned such as 'galaxy photography', 'milky way', etc [10]. In addition, we need a method for handling fuzzy terms used in searching. For example, if we plan to take a vacation and try to find 'cheap hotel' in commercial domain-specific Web search engines such as Hotels.com, Booking.com, etc. At present, they return many hotels irrelevant to user's query 'cheap hotel'. Although they provide hotel list sorted by ascending or descending, they still keyword-based and thus generally return too many results. It is mainly derived from the inappropriateness on reflecting user's search intentions. In other words, they have some problems in handling fuzzy terms. For example, if we want to find a few 'cheap hotel' in Web search engine. In this case, the meaning of fuzzy term 'cheap' depends on individual's subjectivity. To reflect user's search intentions appropriately, we need a Q&A (questions and answers) mechanism. In this respect, we propose an attribute-driven approach to handle fuzzy terms in domain-specific Web search engines. It provides users with a personalized search results in domain-specific Web search engines.

The paper is organized as follows. In the next section, we propose an attribute-driven approach to handle fuzzy terms in domain-specific Web search engines. In Section Ⅲ, to handle the compound fuzzy query with intersection, union connective(s) between fuzzy terms and negation operator, some examples are presented. In Section Ⅳ, a new way of the personalized search based on the values of attributes for fuzzy terms in domain-specific Web search engines is described. We call it perception personalization. In Section Ⅴ, we briefly summarize the differences between proposed method and existing methods on fuzzy terms in domain-specific Web search engines. Finally, we conclude our paper in Section Ⅵ.

II. ATTRIBUTE-DRIVEN APPROACH FOR HANDLING FUZZY TERMS

An advantage of fuzzy set-based modelling is that it is mainly qualitative in nature. Indeed, in many cases, it is enough to use an ordinal scale for the membership degrees. This also facilitates the elicitation of (user/context dependent) membership functions for which it is enough, in practice, to identify the elements that totally belong and those which do not belong at all to the fuzzy set. Fuzzy set membership functions are convenient tools for modelling user's preference profiles. Fuzzy queries are often motivated by the expression of preferences or tolerance and of relative levels of importance [3]. Compared to the existing keyword-based query, fuzzy query provides a better representation of users' preferences and the necessary information for rank-ordering the answers according to the degree to which they satisfy the query. In this paper, we propose Q&A mechanism for narrowing search results by using the values of attributes for fuzzy terms in the phase 2 of Fig. 1

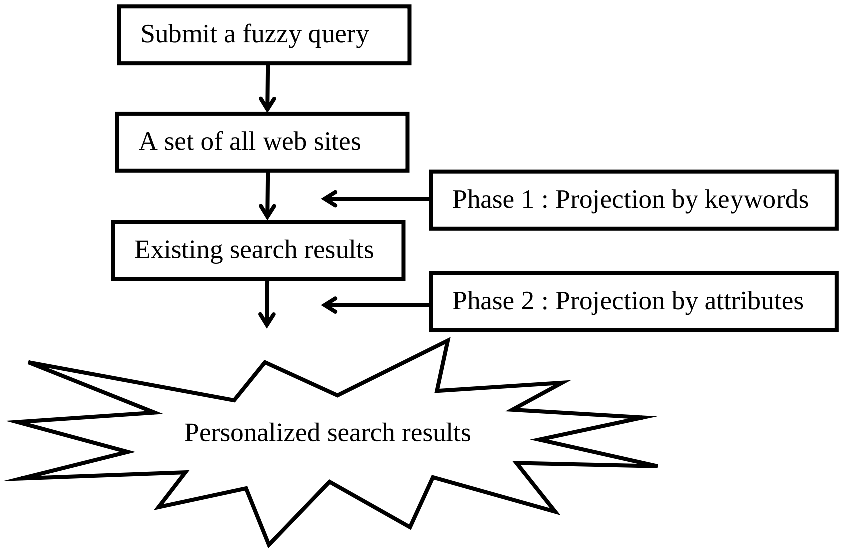
Submit a fuzzy query
A set of all web sites
Phase 1 : Projection by keywords
Existing search results
Phase 2 : Projection by attributes
Personalized search results

Fig.1: Personalized search results in domain-specific Web search engines

Using the attributes on fuzzy terms, we can process fuzzy terms depending on individual's subjectivity. In the sequel, Internet users can narrow thousands of hits irrelevant to user's search intentions to the few that users really want. Although domain-specific Web search engines focus on a specific segment of online content and handle Web pages that are relevant to a pre-defined topic or set of topics, handling fuzzy query in domain-specific Web search engines is still one of the most difficult problems. To reduce the complexity and the degree of freedom of fuzzy query in domain-specific Web search engines, we propose attribute-driven approach for handling fuzzy terms. In the proposed approach, domain-specific Web search engines suggest attributes for handling fuzzy terms. Then user may select their proper values to reflect user's search intentions. For example, if we try to find 'cheap hotel', domain-specific Web search engines recommend attributes including 'room rate', 'no. of visitors/day', etc. for processing the fuzzy term 'cheap'. A user selects the values of attributes that a user really wants. In the computational theory of perceptions (CPT) [7, 8], words play the role of labels of perceptions and, more generally, perceptions are expressed as propositions in a natural language. Computing with words (CW) techniques are employed to translate propositions expressed in a natural language into what is called the generalized constraint language (GCL). In this language, the meaning of a proposition is expressed as a generalized constraint, X isr R, where X is the constrained variable, R is the constraining relation and isr is a variable copula in which r is a discrete variable whose value defines the way in which R constrains X [8]. Attributes for processing fuzzy terms may be a degree of price, distance, size, weight, color, etc., on keywords. They are used to represent a degree of preference, tolerance, importance, etc., on subjectivity in reflecting user's search intentions. Fuzzy terms in fuzzy query can be generally expressed as point value, interval value, multiple values, etc., as follows.

Example 1 (Point value). The fuzzy term 'popular' in searching 'popular national parks' may be specified by using a point value on 'no. of visitors/year'. For example, it may be expressed as a point 3.5 (i.e., 'no. of visitors/year' ≥ 3.5 million)

Example 2 (Interval value). The fuzzy term 'moderate' in searching 'moderate distance' may be specified by using an interval value. For example, it may be expressed as an interval [50, 100] in miles (i.e., distance = [50, 100]).

Example 3 (Multiple values). A veristic variable which can be assigned two or more values in its universe simultaneously will be specified as multiple values. Let U be the universe of natural languages and let X denote the fluency of an individual in English, French and Italian. Then, X isv (1.0 English + 0.8 French + 0.6 Italian) means that the degrees of fluency of X in English, French and Italian are 1.0, 0.8 and 0.6, respectively [7].

III. HANDLING COMPOUND FUZZY QUERY

The compound fuzzy query includes intersection ('*and*' operation) or union ('*or*' operation) connective(s) between fuzzy terms, or negation ('*not*' operation).

Let the set of 'popular national parks in San Francisco' be P = {P_1, P_2, …, P_9, P_{10}}, 'national parks that is located moderate distance from San Francisco' be M = {M_1, M_2, M_3}, and we assume that each P_i, (i=1, 2, …,10) and M_j, (j=1, 2, 3) have their own URL.

Example 4. Consider a fuzzy query that finds 'national parks that popular and moderate distance from San Francisco'. In this query, keyword = {national parks, San Francisco}, fuzzy term = {popular, moderate}, stop word = {that, from}, operator = {*and*}. In this case, attributes may be no. of visitors/year and distance from San Francisco. By applying intersection operator on the respective search result sets of fuzzy terms 'popular' and 'moderate', narrowing search results can be obtained by using the intersection of P and M (i.e., $P \cap M$). In this case, we note that the number of search results become less than 3.

Example 5. Consider a fuzzy query that finds 'national parks that popular or moderate distance from San Francisco'. In this query, keyword = {national parks, San Francisco}, fuzzy term = {popular, moderate}, stop word = {that, from}, operator = {*or*}. In this case, attributes may be no. of visitors/year and distance from San Francisco. By applying union operator on the respective search result sets of fuzzy terms 'popular' and 'moderate', search results can be obtained by using the union of P and M (i.e., $P \cup M$). In this case, we note that the number of search results become 13.

Example 6. Consider a fuzzy query that finds 'not popular national parks in San Francisco'. In this case, attributes may be no. of visitors/year, and logical operator is 'not'(negation). By applying negation operator on the search result sets of fuzzy terms 'popular', search results can be properly obtained. In this query, keyword = {national parks, San Francisco}, fuzzy term = {popular}, stop word = {in, the}, operator = {*not*}. In the case of Example 1, on the fuzzy query 'not popular national parks in San Francisco', proposed search engine returns search results based on the value of 'no. of visitors/year'<3.5 millions.

IV. PERCEPTION PERSONALIZATION BASED ON THE VALUES OF ATTRIBUTES

Internet users do not have to watch all the irrelevant noise. In this respect, personalization technology gives the user a more tailored site. It is the provision to the individual of tailored services, advertisements, products or information relating to products or services.

Personalization of Web pages can be accomplished in numerous ways [2]. Some approaches require the user's participation (typically through filling out a form or questionnaire). Other approaches operate behind the scenes, without depending on user input, for example, by using Web cookies for tracking what user likes to view or by analyzing Web server log for the access pattern and statistics, etc. Search engine logs provide a wealth of information that machine learning techniques can harness to improve search quality. A community-based personalizing Web search approach can leverage the latent knowledge created with in communities by recording users' search activities (i.e., the queries they submit and results they select) at the community level. They can use this data to build a relevance model that guides the promotion of community-relevant results. Google's Web history uses more detail private search activity. Web history can tell you about these and other interesting trends in your Web activity. For example, which sites do you visit frequently? How many searches did you do between 11 a.m. and 2 p.m.? An important problem relating to personalization concerns understanding how a machine can help an individual user via suggesting recommendations [1]. The proposed approach to handle fuzzy terms by using the values of attributes may be considered as a solution relating to personalization for 'understanding how a machine can help an individual user via suggesting recommendations'. It is the perception personalization based on the values of attributes for fuzzy terms. In this respect, it is a new way of the personalized search based on the values of attributes for fuzzy terms in domain-specific Web search engines. Together with existing personalization methods, the perception personalization based on the values of attributes can be used as a complementary and synergistic rather than competitive.

V. COMPARISONS

The central concept of information retrieval is the notion of relevance [6]. A user with a given query for information tries to find any specific results that he/she really wants. There are several models for specifying the representations used for the documents and the queries, as well as the matching of these representations [4]. The most used model is that of the Boolean query based on set theory. Documents are represented as sets of terms and queries are Boolean expressions on terms. The retrieval mechanism does an exact match by classifying documents that satisfy the Boolean query as being relevant, all other documents as being irrelevant. This model is used by virtually all commercial textual-document retrieval systems. However, it is difficult to overcome the limitations of this model such as the inability to handle properly imprecision and subjectivity. The second model is the vector space model where documents and queries are represented as vectors in the space of all possible index terms. The document vectors consist of weights based on term frequencies in the collection, while the query vectors are binary vectors on the terms. The matching is based on a similarity measure between the documents and the query (often involving the cosine of the angle between the query vector and a given document vector). To date, this model leads the others in terms of performance. The third model is the probabilistic model where documents are represented as binary vectors. The queries are vectors of terms with weights based on the estimated probability of relevance of documents with those terms. Like the vector space model, the key advantage is the ability to rank documents on the likelihood of relevance. The fourth model is the generalized Boolean model, where fuzzy set theory allows the extension of the classical Boolean model to incorporate weights and partial matches, and adding the idea of document ranking [6]. Handling fuzzy query in domain-specific Web search engines is one of the most difficult problems. It is mainly derived from the complexity and the degree of freedom of natural language. To reduce the complexity and the degree of freedom of fuzzy query in domain-specific Web search engines, we propose attribute-driven approach for handling fuzzy terms. The proposed search mechanism in Fig. 1 can be explained by SQL-like language: SELECT * FROM {a set of intermediate URLs that satisfies keyword in the document index} [WHERE the value(s) of attribute(s) for handling fuzzy terms are satisfied by the user]. We note that existing Web search engines tend to ignore the importance of [WHERE] part. In the proposed approach, the value(s) of attribute(s) for handling fuzzy terms may be regarded as a constraint on the keywords in document index. Thus, personalized Web search results can be obtained by using the values of attributes for handling fuzzy terns as shown in Fig. 1. We note that user's search intentions can be explicitly reflected by using the values of attributes for fuzzy terms in the domain-specific Web search engines. Thus, the proposed approach provides users with the personalized Web search based on user's search intentions (i.e., keywords, preferences, etc). In this respect, proposed method is a new way of the personalized search based on the values of attributes for fuzzy terms in domain-specific Web search engines.

We briefly summarize the differences between the proposed method and existing methods in Table 1.

Table 1. Comparisons

	Proposed Method	Existing Methods
Search mechanism	One phase using keywords	Two phases using keywords and Q/A with values of attributes
Search results	More personalized	Limited
Mechanism for handling fuzzy terms	Yes	Limited
Personalization method	Values of attributes for fuzzy terms	Web cookies, Web server log, etc.
−For the better personalization, both personalization methods are complementary rather than competitive.		

VI. CONCLUSION

Handling fuzzy query in domain-specific Web search engines is one of the most difficult problems. It is mainly derived from the complexity and the degree of freedom of natural language. To reduce the complexity and the degree of freedom of fuzzy query in domain-specific Web search engines, we propose attribute-driven approach for handling fuzzy terms. As aforementioned, it makes personalized search results by reflecting user's search intentions appropriately in domain-specific Web search engines. Future works may be extended in several directions based on the current proposal : First, more sophisticate method for handling multiple fuzzy terms in a fuzzy query could be developed. For example, consider a fuzzy query that finds 'famous and low-price' hotel. In this case, a weighting method for handling multiple fuzzy terms could be developed. Second, a page ranking algorithm for fuzzy query in domain-specific Web search engines could be developed.

VII. ACKNOWLEDGEMENTS

The author would like to thank prof. L. A. Zadeh, University of California, Berkeley, for his inspirational address on the perceptual aspects of humans in the BISC seminars and group meetings.

REFERENCES

[1] Belkin, N. J., Helping people find what they don't know, Communications of the ACM, 43(8), 58-61, 2000.

[2] Choi, D. Y., Toward a Trend-based Web Page Rank by using Big Data on Smartphones, Advanced Science Letters, Vol. 20, No 10-12, 2134-2137, 2014.

[3] Dubois, D., Prade, H. and Sedes, F., Fuzzy logic techniques in multimedia database querying: A preliminary investigation of the potentials, IEEE trans. on knowledge and data engineering 13(3), 383-392, 2001.

[4] Kraft, D. H and Petry, F. E., Fuzzy Information systems: managing uncertainty in databases and information retrieval systems, Fuzzy sets and systems 90(2), 183-191, 1997.

[5] Battelle, J, The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed Our Culture. Portfolio Trade, 2006.

[6] Salton, S., Automatic text processing: the transformation, analysis and retrieval of information by computer, Addison-Wesley, Reading, MA, 1989.

[7] Zadeh, L. A., From computing with numbers to computing with words – From manipulation of measurements to manipulation of perceptions, IEEE trans. on circuit and systems 45(1), 105-119, 1999.

[8] Zadeh, L. A., A new direction in AI – Toward a computational theory of perceptions, AI Magazine 22(1), 73-84, 2001.

[9] Rao, Leena. "Data-Driven Comparison Shopping Platform FindTheBest Raises $11M From New World, Kleiner Perkins And Others". Http://www. samachar.com/data-driven-comparison-shopping-platform-findthebest-raises -11m-from-new-world-kleiner-perkins-and-others-ndfvNQhejcc.html 2013.

[10] http://whatis.techtarget.com/definition/fuzzy-search.

Dae Young Choi received the B.S., M.S., and Ph.D. degrees in computer sciences from Sogang University, Seoul, South Korea, in 1985, 1992, and 1996, respectively. He was a Research Fellow at the Korea Institute for Defense Analyses (KIDA), Seoul, from 1985 to 1990. He received a postdoctoral fellowship from the Korea Science and Engineering Foundation (KOSEF) in 2000. He was with the Berkeley Initiative in Soft Computing (BISC) Group, Dept. of EECS, UC, Berkeley, as a Visiting Scholar, in 2001. He was also selected as a scholar to be sent abroad under the Professor Dispatching Scheme by the Korea Research Foundation (KRF) in 2004. He was with the Dept. of CSE, Univ. of Colorado at Denver as a Visiting Professor in 2004–2006. Dr. Choi was listed in Marquis Who's Who in Science and Engineering in 2003 and in Marquis Who's Who in the World (2002-2016). He is a Prof. in the Dept. of MIS, Yuhan Univ., Puchon, South Korea. He has a national certificate of Professional Engineer for information processing systems. He was an Associate Editor of the Transactions of the Korea Information Processing Society (Part B) and was on the Editorial Board of the International Journal of Information Processing Systems (IJIPS). His research interests include mobile Web search engines, business intelligence, fuzzy systems, group decision support system (GDSS) and Web personalization.