Web Search: the Protoform approach

Chris Tseng[1, 2] Toan Vu[1]

[1] CS department, San Jose State University, One Washington Square, San Jose, CA 95192, USA
[2] tseng@cs.sjsu.edu (corresponding author), Tel: 1-408-924-7255, Fax: 1-717-754-6112

Abstract

We investigate how web queries in Protoform semantic can be used for perception-based fuzzy semantic search. Web queries are first represented in Protoform. Iterative searches are performed on the Protoform variables to extract a subset matching the semantic description of the search queries. In the last stage of this perception-based search process, we show how search results can be sorted in accordance to their relevance with Fuzzy Semantic Search engine.

Keywords: Protoform. Semantic Search. Fuzzy Logic. Search Engine.

1. Introduction

Existing popular web search engines such as Google can perform keyword type of stateless search. The search engine can hardly handle an efficient search when complex phrases or complete sentences are given as search inputs. Much of this deficiency has something to do with the lack of capability in perceiving the search input semantically. The GCL (Generalized Constraint Language) and PTL (Protoform Language), both proposed by Zadeh [1,2], are ideal ways to abstract and precisiate search inputs semantically for a more efficient web search process.

Fuzzy Semantic Search (FSS) can leverage on the semantic fuzzy synonyms of search keywords to expand the search for a more relevant web search [3]. Once the input search statements are transformed into PTL format, one need to map the description in PTL to a form acceptable by the existing keyword-type search engine, e.g., Google [4], for the physical search query. We will show in this paper how this can be done with FSS engine.

We begin by presenting an overview on FSS and the semantic structure of PTL in section 2. In section 3, we provide an example to illustrate the steps involved in processing a search statements with PTL followed by FSS. Future research and related issues are described in the conclusion in section 6.

2. Protoform and Fuzzy Semantic Search

Fuzzy Semantic Search (FSS) is an innovative perception-based search methodology that has shown considerable improvement over conventional search methods in obtaining more relevance results [3]. FSS has been shown to be much more effective in retrieving web pages relevant to the search query as compared with commercial search engines such as Google and Yahoo. The improvement is even more impressive if the scheme is applied to web sites with matching semantic content included in the FSS design.

FSS, as shown in Figure 2.1, takes advantage of the search query semantic and expand the search to include information related to the fuzzy linguistic, FLS, and numeric, FNS, variables [5] of the query input. This search methodology has three stages in it process. The first step is to take the main subject of the search keywords and send them to search engines to obtain relevant web pages. In the second stage, the system looks for web pages returned from stage 1 that have matching fuzzy semantic counterparts related to the keyword constraining the main search keyword. Finally, we deduce and rank the web pages in accordance to their overall matching membership values in the fuzzy logic sense.

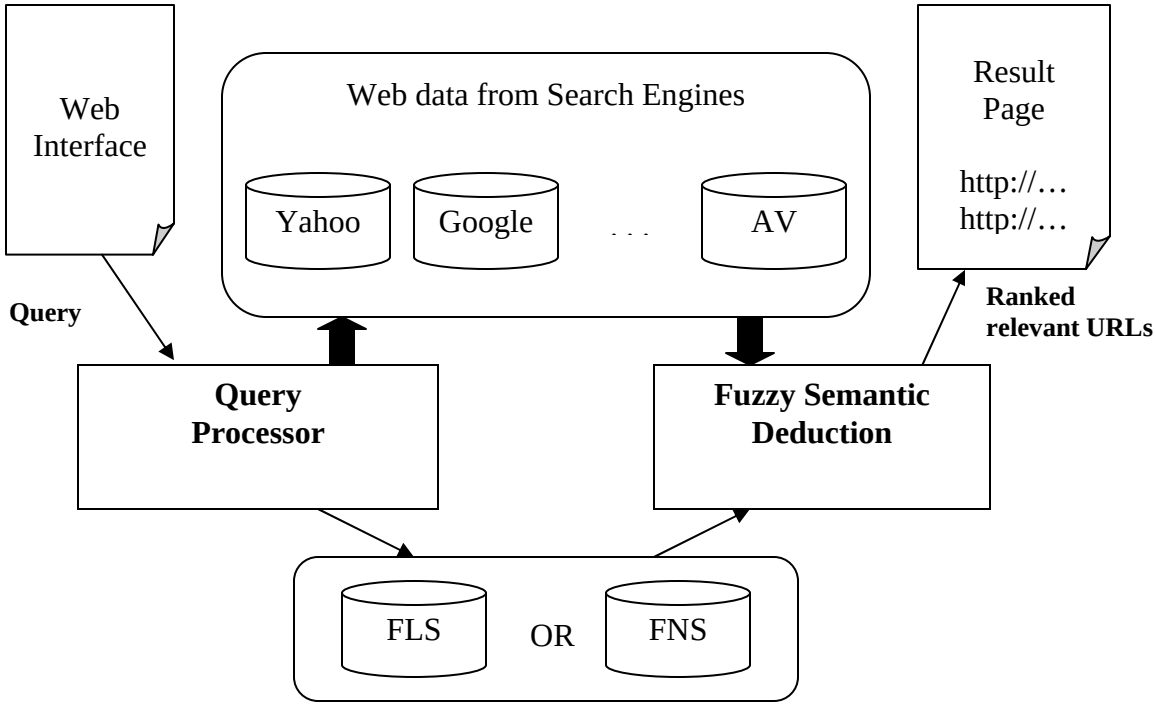

Figure 2.1: Perception-based Fuzzy Semantic Search

FSS can take statements in GCL (Generalized Constraint Language) or PTL (ProToform Language) form and perform searches once the keywords and the constraint variables are abstracted from the search query. An input statement in natural language can first be converted into the GCL and PTL format. In so doing, the statement is abstracted into variables and the affiliated constraining constraints. In general, a statement in Protoform can be expressed as

$$V_1 (V_2 \dots (V_n))) \textit{ is } C_1, C_{1,\dots} C_m, \quad (2.1)$$

where Vi, i=1,…n, are the variable and Ci, i=1,…m, are the constraints constraining the variables. In subsequent integration with FSS, we treat all the variables V_i as subject keywords and constraints C_j as fuzzy variables constraining the subject keywords. Keyword searches are performed on $V_{i,}$ i=1,…n, progressively to narrow down the matching web pages that contain matching subject keywords. The semantic variable V_i has higher priority of operation over the outside ones, V_j, i < j. Thus, the apparent hierarchical relationship between the V_i in PTL form, V_k has higher priority or more relevance to the topic then V_j where k>j, indicates that the process should start with the search from the inner most variable, V_n, and recursively continue with the process applying to the outside variables. This process essentially starts with a set of web pages considered to contain the subject V_n and continue to look for subsets that contain subjects like $V_{i,}$ i= 1,…, n-1 parsing through content in the webpage subset. This type of multi-pass search is known to be a powerful alternative in search engine such as Copernic [6]. Once we reach the subset that contains all the subjects, we then apply all the constraints Cs, i= 1,…, m on the webpage subset containing the subjects V_1 to V_n.

FSS can use its own search engine or augment itself as an add-on layer to the commercial search engine like Google to continue the web search process and produce the output. Fuzzy membership function score and distance rank score are applied in FSS and produce a ranked output according to its relevancy to the search query [3]. The overall architecture is described in Figure 2.2.

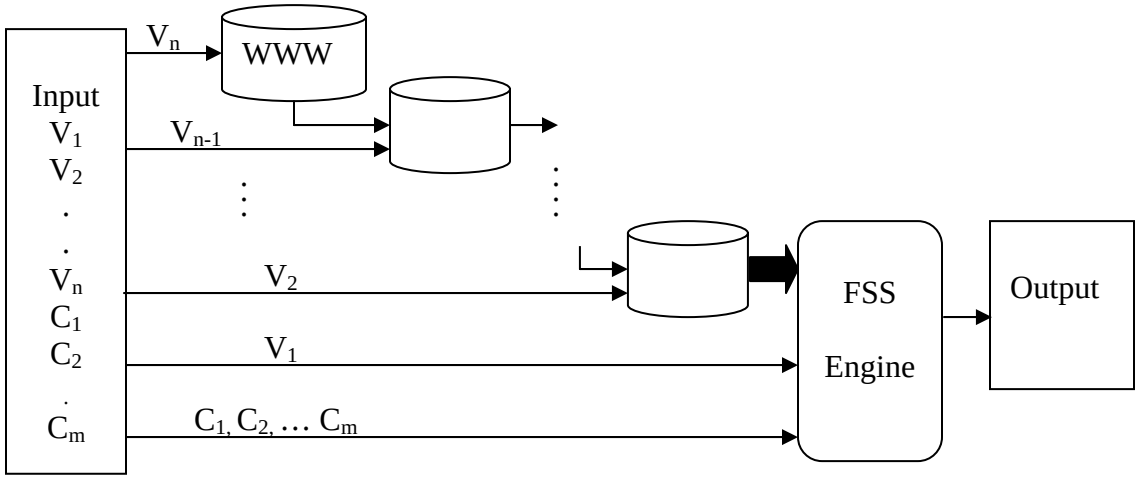

Figure 2.2: Fuzzy Semantic Search with input in Protoform form

3. Example and case study

Consider the statement
"***Show me warm weather cities***." (3.1)

Figure 3.1 displays the overview of how the statement (3.1) is processed by our proposed methodology. In the initial phase, a statement is translated into its appropriate GCL form. This example statement can be represent by the equality constraint of the GCL form

X is R, (3.2)

where X is weather(city) and R is warm.

Abstracting the above GCL form to PTL form, we have

A(B) *is* C, (3.3)

where A is weather, B is city, and C is warm.

The second phase searches the keywords one by one starting with V_n from the world-wide-web and continues filtering the result with V_i until reaching V_1. In this particular example, where n = i = 2. We only need to search for all web pages with the keyword ***city*** from the web. No further filtering of the result is required.

The last phase makes use of the FSS engine that was developed in earlier research [3]. We applied *warm* and its fuzzy counterparts to *weather* in the FSS engine over the list of web pages obtained from the city search result in the second phase.

The results are then ranked and displayed according to its membership function score and distance score evaluated by the FSS Engine. The output will have the cities with warm weather on the top of the list, follow by cities with closely related description such as mild weather or temperate weather, and at the bottom will be cities that have hot weather or

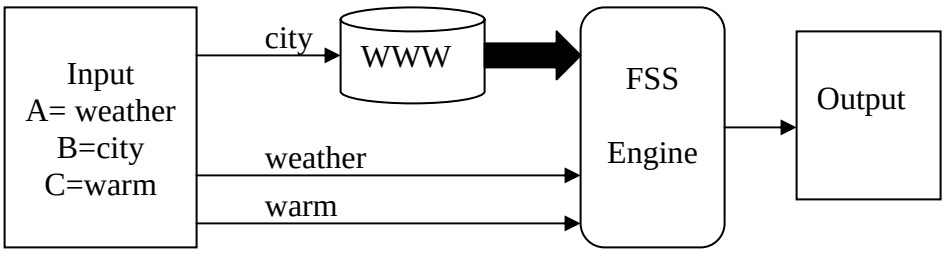

Figure 3.1: Example on protoform input to web search

cold weather that are semantically distant from the constraint keyword, ***warm***.

4. Conclusion

The issue of how search engine can understand complete and complex query statements with Protoform semantic representation and Fuzzy Semantic Search is discussed in this paper. Protoform is shown to be suitable for representing search query in a form that clearly separate search subjects and their affiliated constraints. Fuzzy Semantic Search Engine is an efficient tool for extracting web pages matching the search subjects with the constraints implied in the original search query statements.

Numerous important subjects await more research results in this direction. How a natural language statement can be mapped to a Protoform statement is certainly one of the most important issues to be worked upon. A question-answer type web search can truly be realized in the near future with this approach

5. References

[1] L. Zadeh, From computing with numbers to computing with words. From manipulation of measurements to manipulation of perceptions, Ann NY Acad Sci, 929, 221-252 (2001).

[2] L. Zadeh, Precisiated Natural Language, American Association of Artificial Intelligence, 74-91 (2004).

[3] C. Tseng, T. Vu, "A perception-based web search with fuzzy semantic", *Fuzzy Logic in Semantic Web, Elsevier*, 2005

[4]"Google online search", *http://www.google.com*.

[5] L. Zadeh, "Knowledge representation in fuzzy logic", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 1, No. 1, March 1989 pp.89-100.

[6]"Copernic search engine", *http://www.copernic.com.*