Research on Intelligent Semantic Search Based on HowNet

Feng Wang¹*, Hui Zhang², Yizhen Wang³ and Guanghua Zhang²

¹School of Mathematics and Statistics Science, Ludong University, Yantai 264025, China, and Key Laboratory of Language Resource Development and Application of Shandong Province; Yantai 264025, China; ²College of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, China; ³Tilburg University, Tilburg, the Netherlands

Abstract: HDWiki is similar to the famous Wikipedia but as being a major encyclopedia in Chinese, it includes vast investment of manual effort and judgment. This structured information is divided into three categories: concept, relationship and entity. This paper describes a novel symbol oriented search that facilitates the user to format unambiguous query, especially to format complex logical query which needs several queries to get answer on traditional search but once incorporated here, it builds a bridge between what the user knows and what the user wishes and between the intent of user query and the understanding of search engine. This facilitation and understanding based on the ternary relationships are everywhere in HDWiki and HowNet powered semantic similarity measure for knowledge matching. A logical semantic oriented search system is realized and a detailed user study is conducted with users and it is found that this logical semantic oriented search and its underlying knowledge base provide significant advantages over conventional search. It offers assistance to almost every query especially complex logical queries, making the query entry more clear and efficient for user as well as for search engine, broadening the processing scope of complex logical query in the search, improving the relevance of the searched documents and increasing the probability of achieving what the user wishes to get.

Keywords: Concept, HDWiki, HowNet, Information retrieval, Logical symbol, Query formulation, Semantic similarity, Ternary relationship, Wikipedia.

1. INTRODUCTION

Whenever there is a need to acquire new knowledge and people resort to the ubiquitous search engines, the struggle is always with the same fundamental paradox: how can one describe the unknown? This diachronic question still relevant on today’s internet era is precisely what must be done to form a query [1]. What makes matters worse is that search engines have insufficient capacity to comprehend and deal with these logical descriptions as people do; they instead treat a query as nothing more than an excerpt: a few words or phrases -- from a relevant document. The search engines are likely to provide good relevant documents only when a query contains multiple topic-specific keywords that accurately describe the information a user needs. In short, in order to find or acquire relevant information, one must already know a great deal of what is being sought.

For knowledge seekers, there are ravines between what they know and what they wish to know, between their vague initial query and the concrete concepts and entities available, between the intent of the query and the direction to which the search engine leads to. One possible bridge to what knowledge seekers need is provision of several semantic logics and an encyclopedia: a map of semantic relations between terms and phrases.

Knowledge seekers who cannot identify what terms could impel the query to be clear and concrete as concepts even have not realized that the ambiguity of their query could benefit from an encyclopedia that automatically ensures the specific concepts by the logical inferring based on its terminology of relationships, concepts and entities, together with recommending a list of the most possible concepts to them. Those who cannot formulate a specific query, especially a complex logical query also could use the structured knowledge of an encyclopedia and semantic logical syntax to straighten out the terms to make it precise for them as well as for the search engine. This could have the potential to greatly advance the art of information retrieval. There are freely available large encyclopedias like Wikipedia, HDWiki (http://www.baike.com/), etc., and this research is based on HDWiki, the largest Chinese online encyclopedia like Wikipedia.

HDWiki articles contain, besides free text, various types of structured information in the form of wiki markup. These various kinds of information can be divided into three categories: concepts, relationships and entity. With this vast knowledge base, a way should be found to make the knowledge structured to be used precisely and in order to make searching more facile. The next section describes the HDWiki knowledge and its extraction way, section 3 presents our methodology to construct a bridge -- our logical semantic search process and section 4 introduces the semantic similarity measure powered by HowNet-- the indispensable concatenation at the end of the bridge. In section 5, the search system is built.
and the experiments are analyzed and the context of research surrounding this work is described in section 6. The paper concluded with a summary and the pointing of issues needed to be addressed in future research (section 7).

2. HDWiki KNOWLEDGE

Similar to Wikipedia, HDWiki contains a vast investment of manual effort and judgment and it is an open, constantly evolving encyclopedia. In a general circumstance that binary relations are present everywhere even in one article, they are included at both sides of a colon or between subheads and the below text content. Furthermore, the title is relative to all the binary relations in an article. These can be regarded as ternary relationships and each of them contains the article title.

The title of an HDWiki article is defined as a concept. A binary relation abstracted has the left relationship being a ternary relationship, and the other one which is also the last one of a ternary relation can be entities or concepts. So HDWiki knowledge is divided into three categories: concept, relationship and entity.

1. Entity is usually composed of some terms; it can be a concrete human name, an organization, a book, a game, a place, a concrete event, etc. You cannot make sure who he is just according to a name, because there are many people. Similarly, it cannot be determined what something is according to an entity. In addition, entity could contain a few concepts (defined below).

2. Concepts stand for definitive information. For example, there is one “王菲” who is China's top diva, just like there is only one article about this “王菲” in HDWiki.

3. Concepts or entities are associated with the relationships of this concept. For example, the relationships of China's top diva “王菲” include the blood type, the relations of characters, musical works, anecdotes related to her, etc.

The logical semantic oriented search relies on - HDWiki knowledge base. This information should be made practicable and available on the web under an open license. The DBpedia extraction framework adopted from Wikipedia is quite good [2], but the knowledge it does not cover Chinese language. According to the regular DOM architecture of HDWiki web, the ternary relationships are abstracted through THML parsing and this can be carried out in real time.

3. LOGICAL SEMANTIC ORIENTED SEARCH

How can the HDWiki knowledge base be used to facilitate search? - What methodology do we take to build the bridge for the user and the search engine? Here, five logical symbols are defined to guide both the user and the search engine. These are not the existing logical operators (AND, OR, NOT) but can be used together. Guide symbols described are “#C#”, “#E#”, “#P#”, “#W#” and “#S#”. The two previous symbols are used to ascertain the concept of an entity. “#C#” can be used with concept words directly, or “#E#” can be employed with specific entity or relationship which can then logically deduce the concrete concept based on the ternary relationship knowledge. Among concept, relationship and entity, the third element can be obtained if the two elements of this ternary relationship are already present. If there is a concept, then the entities can be obtained corresponding to a special relationship with “#P#”. And any third element using “#E#” can be obtained with two other elements. However, there may be a fundamental factor which cannot be ignored and then the question arises as how the terms can be matched which the user inputted to the element in the ternary relationship; this is provided in detail in section 4. Though it is possible that the above search based on ternary relationship can be carried out through the general search but there is a blurry logic that could not be comprehended by this. For example, the resemble logic. Here, a symbol “#S#” is introduced which is used to deal with those resembling logics with the help of semantic similarity measure.

Generally, the first step is to ascertain the concepts of user query to clear the search subject and then it can be helped to straighten out the search ideas for some complex logical queries. All of these are relevant for both the users and the search engines as the ravines are between them, and both of them should strive.

3.1. Semantic Concept Oriented Search

With respect to the term based query, binding a special semantic concept to a term can help to deal with word sense disambiguation problem in the information retrieval. For example, “王菲” has three concrete semantic concepts: “中国女歌手”, “四川省广元市代市口”, “大同大学教授” in HDWiki articles, and it cannot be made sure that who is “王菲” just according to these two characters.

Here, a binding symbol “#C#” is defined; the format is:

Entity#C#Concept

In the formulation above, “Entity” means the term with a vague search target and “Concept” means a concrete and unique semantic concept, abstracted from HDWiki.

For binding a semantic concept to an entity, the first direct way is binding the concept words of HDWiki, for example: “王菲#C#中国女歌手”。But most of the time, users do not know what are the special concept word, for example: “王菲” is a “歌手”, and she is a “女歌手”, she also is a “歌后天后” and there would be many more honors for her. So it is difficult that the concept terms bounded exactly match to the concept of HDWiki; this is a serious mismatch problem.

According to the semantic similarity between concept words what the user binds and HDWiki concepts, selecting the maximum semantic similarity as the right semantic concept would be a good solution to the mismatch problem mentioned above. The semantic similarity considers the co-occurrence and semantic correlation of the words; its detail is described in section 4.

In addition, it is known that a concept is always correlated to corresponding entities with some relationships. It means that if a concepts list is obtained, and with a special relationship and its corresponding entity, the concept can be ascertained which has this ternary relationship. Using the same symbol above, with the format: Entity#C# (Relationship Entity).
For example, query “王菲#C#(女儿李嫣)” means what user required. “王菲” is a concept that she or he may have a “女儿” named “李嫣”。The simple way to get this concept is to match “李嫣” to “女儿” of all concepts, and the concept of this can be found “王菲” as “中国女歌手”。

The logic shown above also needs semantic similarity calculation during the matching process, which is feasible, but two match processes should be executed; the first for relationship and then the entity should be matched in the second one. In consideration of the unnecessary process procedure, the entity of each concept can be straightforwardly checked using a new binding symbol “#E#”:Entity#E#Entity.

With this symbol, we can bind “Entity” (the first one above) with entity or relationship (the second “Entity”) for the semantic concept binding. For example, the second “Entity” can be a relationship (e.g. “□影”) or an entity (e.g. “李嫣”). In other words, “#E#” means the concept has a specific entity or a specific relationship.

In a nutshell, on the basis of HDWiki, with binding symbol “#C#” and “#E#”, the concept can be ascertained which the user required according to the bound concept or relationship or entity.

The interaction with users can strengthen the performance of semantic concept binding and this can make the semantic similarity calculation matching more better. It is that the concepts what the terms matching process procreate except the decided one can be recommended to users just in case; the semantic similarity measure process can narrow the selection scope for users to enhance the efficiency and convenience for them. A concise and visualized recommendation widget is designed as shown in section 5.

3.2. Logical Semantic Search

In many cases, what users know is just related or similar to what they wish to know. But mostly, the traditional search engines could not deal with this “related or similar” logical. And this may be worse when the dealing is with any complex logical query. The user may have to make judgments in many search processes and even have no idea to get what they wish through the search engine. These are the issues which can be solved below by orienting the power of HDWiki knowledge base.

3.2.1. Combined Logical Search

If a user wants to know all the movies of a star, the existing powerful search engine may show a list of movies of this star and may be some other information related to him. For example, inputting “古天□□影”, Google would show the movie list and brief information of “古天□”。Now, the user needs to search an actress who is the heroine in a film of star “古天□” and does not know the film name and just thinks that this name is similar to “蜜糖”。The query what the user inputs is “古天□□影名似蜜糖女主角”, and the content (Fig. 3) what the search engine returns with contains the query terms, which has nothing to do with what the user wished to find.

In order to find the actress, the user would have to keep trying various queries by paying a lot of time and patience. Finally, the answer may be found in a “smart” way that if the user has queried for the movies list of “古天□”, then the user would have to check them one by one according to “蜜糖” spending a lot of time, then the user may have to try the second query with a probable movie, then the third query, … and then may find the correct film name “蜜糖蜜□”.

Now, it is known that HDWiki has movies list of “古天□”。The semantic similarities between each movie and “蜜糖” can be computed to get the most probably right film name, and at the meantime, some other movies can be recommended to user with high semantic similarity. This seems very feasible. What needs to be done is to make the search engine understand the logical operation: Get the movies list first, and then calculate all the semantic similarities between movies and “蜜糖”, select the maximal one to take the place of “蜜糖” in the query, and display other film names which have high semantic similarity as substitution.

Here, two binding symbols “#P#” and “#S#” are defined to guide the logical operation. They are in the following formats:

- Concept#P#Relationship
- Entities#S#Terms

Symbol “#P#” refers to getting all corresponding entities in the “Relationship” of “Concept”. E.g. “古天□#P#□影” means to get all the movies of “古天□”.

“#S#” means selecting an entity which has the maximum semantic similarity with “Terms” from “Entities”. E.g. find a movie which has the maximum semantic similarity with “蜜糖” from the cinematographic work of “古天□”, the movie “蜜糖蜜□” will be found.

With the ternary relationships of HDWiki structured knowledge and combining the two symbols, the query “古天□#P#□影#S#蜜糖女主角” is processed to be the new query “古天□□影蜜糖女主角”。And the right web results can be obtained even through sending this processed query to Google. This is shown through our practical search system in section 5.

Furthermore, symbols “#P#” and “#S#” can be used separately. For example: Query “王菲#P#女儿” can get “女儿” of “王菲” from HDWiki; And we can get “美式足球” by the query “球□运□#S#足球” from HDWiki.

Without doubt, the users could expeditiously format other various complex logical queries with these logical semantic symbols based on the above concepts, relationships and entities of HDWiki knowledge, as well as format simple queries distinctly.

Just like the recommendation what when we bind concept using the symbol “#C#” do, in case that what symbol “#S#” selected by semantic similarity measuring probably be not user required, the recommendation widget would play a role again.

3.2.2. Query by Example

May be you would run into a situation that you wished to be similar to another you knew but had no direct link to it. For example, you know that Obama is a leader of USA but you cannot make sure that he is the president, and you want to know what the country Putin leads is, besides
Obama and Putin have the same leadership. The symbol "#E#" is defined to get the one left along with knowing other two in a ternary relationship. Syntax format is shown below:

#E#(Concept Entity)

Here, both “Concept” and “Entity” can be any two concepts, relationships and entities in all kinds of left-to-right orders. So, the query can be formatted: Putin #E# (USA Obama). This could be a process to be the new query: Putin president.

3.3. Priority for Processing

The query can be carried out using the above-defined five binding symbols in many compound modules. Each symbol has a different function, so their priority is defined to be dealt with.

Generally, the query with logical operator is processed in the left to right order, and this rule is also applicable to our logical semantic symbols, but the semantic concept should be made sure at the beginning since binding semantic concept symbol is set "#CH#" and "#H#" having the first priority. Query by example can be analyzed anytime during the searching, as well as content of a specific relation can be obtained selecting the most correlated entity. The three "#P#", "#S#" and "#E#" should be dealt with in the general left to right order.

In addition, these logical semantic symbols can be used together with the normal logical operators including AND, OR, NOT. There is no conflict between the two groups operation.

4. SEMANTIC SIMILARITY MEASURE

As stated earlier in section 3, the required information cannot be obtained from HDWiki knowledge just by simply matching the terms. The simple example is that the user cannot get the concept “中国女歌手” through what HDWiki knowledge has by just matching the term “女歌手” to it, and it is more difficult to match “甜言蜜语” with “甜蜜”. This consequently impacts the deployment of logical semantic search based on HDWiki, so what is needed is the indispensible concatenation at the end of the bridge with user terms.

Most often, there may be same characters between the HDWiki terms user wished and the terms user bound, e.g. “女歌手”, and further there are some uniform elements in between their semantic expressions. These two relations are defined in terms of similarity and semantic correlation. What is needed is the similarity expression summed by these both.

It is important to note that the purpose of this semantic similarity calculation is to get the terms which have the most similar meaning with the contrasted terms. And this is different from what many other researches do.

Each of the concepts, relationships and entities consists of several terms which are here considered as phrases. And the semantic similarity is measured between the two phrases.

The similarity of terms mainly represents the influence of the same characters on the two phrases composed by the terms. Further, it should be considered that same terms have larger weight than the same characters. So the similarity of terms is measured as follows:

$$TS(p, p_i) = \alpha \cdot \frac{Sch(p, p_i)}{Lch(p)} + \beta \cdot \frac{Stel(p, p_i)}{Lte(p)}, \quad \alpha + \beta = 1 \quad (1)$$

Where

Phrase $p$ is the contrasted phrase and $p_i$ is one of the selection sets,

$Sch(p, p_i)$ is the number of the same characters between two phrases $p$ and $p_i$,

$Lch(p)$ is the length of phrase $p$ in character,

$Stel(p, p_i)$ is the number of same terms between the two phrases,

$Lte(p)$ is the length of phrase $p$ in the term,

$\alpha$ and $\beta$ stand for two weights.

General semantic dictionary WordNet simply consists of the concept down to a tree-level system of concepts, but HowNet here is trying to describe every concept with a series of sememes [3, 4] and each term can be expressed by a number of concepts. In this work, the semantic correlation of two phrases is based on HowNet.

First of all, a phrase is divided into a list of terms using the Chinese segmentation system ICTCLAS [5]. Corresponding sememes of each term are obtained from HowNet official provision, that is, each term is defined as a vector $t = [e_1, e_2, e_3 ... e_m]$. $e_i$ is the $i^{th}$ sememe of term $t$ and each sememe in the vector is one dimension. DAI’s [6, 7] modify strategy is adopted for calculating the similarity between two sememes in HowNet:

$$Sim(e_i, e_j) = \frac{\alpha e^{\beta h} - e^{-\beta h}}{d + \alpha e^{\beta h} + e^{-\beta h}} \quad (2)$$

$$Sim(e_i, e_j)$$ is the similarity between sememe $e_i$ and $e_j$. $d$ is the distance between $e_i$ and $e_j$ in the sememe tree, $\alpha$ is a parameter which means the distance when the similarity is 0.5. $h$ is the depth of the first common parent node of the two sememes and $\beta$ is a smoothing factor. $\alpha$ is set to be 1.6 and $\beta$ is set to be 0.16 as defined in [15].

Supposing that the term $t$ has $M$ sememes $t_1, t_2, ... t_M$ and the other term $k$ corresponding to $N$ sememes is $k_1, k_2, ... k_N$, then the similarity between these two terms is:

$$Sim(t, k) = \frac{\sum_{i=1}^{M} \max_j \left( Sim(t_i, k_j) \right)}{M}, \quad j = 1, 2, ... N \quad (3)$$

Assuming there are two phrases $p$ and $v$. $p$ is segmented into $E$ terms $p_1, p_2, ... p_E$ and $v$ is divided into $F$ terms $v_1, v_2, ... v_F$. The similarity between $p_i$ and $v_j$ is $Sim(p_i, v_j)$, where $i = 1, 2, ... E$, $j = 1, 2, ... F$ [8] and the semantic correlation between the two phrases $p$ and $v$ is:

$$SC(p, v) = \sum_{i=1}^{E} \max_j \left( Sim(p_i, v_j) \right), \quad j = 1, 2, ... F \quad (4)$$
Finally, the semantic similarity of the two phrases is:

$$SS(p, v) = \gamma \cdot TS(p, v) + \delta \cdot SC(p, v), \quad \gamma + \delta = 1. \quad (5)$$

Where $\gamma, \delta$ represent weights respectively.

Segmentation system developed by the Chinese Academy of Sciences is used.

The phrases are sorted by the semantic similarity in the descending order; the maximum one is decided to be the right one and the other phrases whose semantic similarity is higher than a specific threshold value can be used as a recommendation for the users.

5. REALIZATION AND ANALYSIS

On the basis of the popular search engine Google, our own logical semantic search system is designed that embeds the processing module of logical semantic symbol based on HDWiki and HowNet. Users can compare the web search results between Google and our engine intuitively. Fig. (1) describes the architecture of this search system.

The concept bound search instance is described in Fig. (2).

And a big difference can be seen between the logical semantic search and general search with the query “古天乐#张柏芝#甜蜜女主角” as shown in Fig. (3).

And a concise and effective recommendation widget is designed like this in Fig. (4).

During the practice of realizing this search system, it was found that the logical semantic oriented process can be carried out before the first submit during the searching process. In other words, when the user finishes one map of binding phrase which is part of the whole query in the input box, this can be processed immediately. This may have greater efficiency and can reduce the second searching to a great extent.

6. RELATED WORK

The central idea of this research is to extract encyclopedic information from HDWiki and to use the structured knowledge to facilitate query formulation and oriented search semi-automatically. The interactive aspect is that the users are required to input the oriented logic with the five symbols initiatly and may also need to do some selecting work according to the recommendation. The automatic aspects involve abstracting of the HDWiki knowledge and processing of the query according to the logical semantic symbol with the semantic similarity measurement based on this structured information.

The most used encyclopedia in the information retrieval research is Wikipedia. The research focuses on the resources including hyperlinks, categories [9-11] and structured knowledge [1, 4]. Some users may wish to make the Wikipedia knowledge more convenient and effective; therein the structured knowledge of Wikipedia can be abstracted using DBpedia extraction framework in order to facilitate the com-

![Fig. (1). Architecture of logical semantic oriented search.](image1)

![Fig. (2). Concept bound search practice.](image2)
plex query of Wikipedia. A new search engine Koru would help guide the user interactively and would expand query automatically based on the structured knowledge, however the user would need relatively more operation time during this reciprocal process.

The work follows this theme but differs in the way that it considers that all the structured knowledge of HDWiki can be used as ternary relationships and that this information can be abstracted according to the HTML structure in real time and what is recommended to users needs less thinking and action.

Word sense disambiguation (WSD) is a natural and well known approach to the vocabulary problem in information retrieval [12] and is traditionally considered an AI-hard problem [13]. Machine learning and statistical techniques to WSD have been proved to have inherent limitations. Actually, many WSD processes are powered by WordNet [14].

Compared with these methods, the concept can be obtained directly from HDWiki with the semantic concept binding symbol, and in the future, some steps can be carried out to get it based on the structured HDWiki knowledge without any symbol.

There are a great deal of similarity measures based on WordNet [15-18]. Most of them concentrate on the similarity between two concepts mapped in the semantic network and emphasize the horizontal and vertical correlations. Yet, the similarity what we focus on is pursuing the same meaning and a simple semantic similarity measuring powered by the HowNet is chosen [19] to resolve it. In addition, phrase semantic similarity could be helpful.

CONCLUSION

This paper introduced a neoteric logical semantic oriented search that makes advantage of the manual structured
knowledge of HDWiki. Our intuition is that the information in this encyclopedia could be abstracted in ternary relationships combined with the semantic similarity measuring powered by HowNet for knowledge matching and these together could provide a lot of assistance in facilitating users to format unambiguous query and complex logical query to get answer even more quickly. This can allow the users to apply the knowledge found in HDWiki to their retrieval process easily, effectively and efficiently. It has been experimented with varied query requirements on a realized search system and it was able to lend assistance to almost all queries especially the complex logical query which needed several queries on traditional search but once obtained, this significantly improved the retrieval performance. Our goal in the future is to reduce the oriented symbols to simplify the formatting as much as possible, try to give the result directly to seekers and to make good use of the concept of HDWiki and its corresponding article to improve the information retrieval in the automatic query expansion.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This research is supported by the National Natural Science Foundation of China (61300120), the University Scientific Research Foundation of Hebei Province of China (YQ2014036, YQ2013032).

REFERENCES


