Meta Search Engine Powered by DBpedia

Boo Vooi Keong
UMS-MIMOS Center of Excellence in Semantic Agents
School of Engineering and Information Technology
University Malaysia Sabah
Kota Kinabalu, Sabah
nogias1685@yahoo.co.uk

Patricia Anthony
UMS-MIMOS Center of Excellence in Semantic Agents
School of Engineering and Information Technology
University Malaysia Sabah
Kota Kinabalu, Sabah
panthony@ums.edu.my

Abstract— The evolution of information retrieval technology on the web has led to the idea of semantic search engine in which it understands the meaning and the context of the search query. As a consequence, the search results returned by this type of search engine should match closely with the query. However, the Web is still dominated by Web 2.0 in which information and data is presented in an unstructured manner and is only fit for human consumption. Hence, building a semantic search engine is a very challenging task and there is still a lot of improvement that needs to be done to achieve the desirable results. As an example, if we search for "food that is not halal", existing semantic search engines still ignore the term of "not" resulting in inaccurate search. In view of this problem, this paper proposes a semantic meta search engine that utilizes the power of a traditional search engine (Google) and enriches the search result using DBpedia as the knowledge base to produce better results. This paper also describes the application of the knowledge base contained in DBpedia to deliver an improved search engine.

Keywords: semantic search, search engine, DBpedia

I. INTRODUCTION

The difficulty of information retrieval on the web is proportional to the size of information available on the web. Based on the current web scale, it will be pointless for a user to search for information on the web by just surfing through the web solely based on the hyperlinks provided by a given webpage since it will take forever to go through all the contents in the World Wide Web. This is the reason why a search engine is very useful in information retrieval. A search engine will go through the contents of the web on behalf of the user and tries to provide all the "answers" based on the known content.

Search engines have evolved over time in stages as summarized in [1] and [2]. In the first stage, when the World Wide Web has just started to grow, search engines only focused on the ability to index the web resources and then answer user query based on keyword search. However, as the web kept growing in the subsequent years, issues on the quality of the search results arose.

There are a few points about search engine that were stressed in [1]:

1. Information on the web is growing so rapidly such that it is impossible for user to maintain a list manually, which covers any popular topic that a user might be interested in.

2. Junk results just tend to fill out the top list of the search result instead of meaningful result without a good ranking algorithm. As they claimed, in November 1997, only one out of four commercial search engines will return their search page in the top ten results in response to their name as a search query.

3. As the web grows over the years, humans are no longer able to keep track of the vast amount of information available on the web.

To solve this problem, they concluded that a powerful search engine that is capable to return a high precision result is required. The emergence of Google has led to a new standard in search engine, the important of quality in searching. Google, with its unique way of ranking pages in the web has become the de facto standard for searching information on the web. Google still has some limitations since it is a searching mechanism that is based on keyword. Both the indexing of web pages and the matching of user query do not consider the meaning within it. As the knowledge or concepts in the world grow over the years the same word could refer to different concepts in different domains. Hence, answering a user query precisely without knowing the meaning of the search query is challenging. The term "cat" refers to more than 10 possible meanings in Wikipedia. This is one of the reasons for the construction of Web 3.0 in which knowledge is defined under the standard, Resource Description Framework (RDF), a standard that is used to show relationship between concepts [3].

Semantic search engine searches based on the meaning of the input query rather than considering a query as a group of string. The results returned from a semantic search engine query are supposed to address the requirements of the user based on the meaning and the context of the query. An example of a semantic search engine is Hakia [2]. Hakia search engines tries to bring search experience as close as possible to the way human interacts in real life. Hakia claims that traditional search engine has problems in ranking based on statistics where less popular topic will not be ranked properly. It also claims that a semantic search must fulfill certain criteria such as:

1. Understand the query similar to human brain.
II. REVIEW OF SEMANTIC SEARCH

There are a significant number of researches on the development of semantic search engines. Some of the notable semantic search engines include Hakia, Koru and Look4. Hakia is an ontological semantic and natural language processing (NLP) based search engine. Hakia processes information using its proprietary core semantic technology called QEDEXing [2]. The idea behind QEDEX (Query Detection and Extraction) is that, instead of indexing all the words into a table for web resource like the traditional search engines, Hakia parses a sentence as a whole. Each sentence is assumed to answer a certain question. In other words, QEDEX decomposes sentences into meaningful knowledge sequences. For example, the sentence “a cat is on a mat” is the answer for a question “where is a cat?” or “what is on the mat?”. Hakia applies fuzzy logic algorithm to parse the entire sentence, generating all possible questions that a page could answer and indexes the page based on the question list. When a search query comes in, it is assumed that it is asking for certain question. Based on the query (question), Hakia will just pick the result from the question list generated previously. The concept of Hakia is shown in Figure 1.

Koru is another semantic search engine that claims to be able to "harness Wikipedia to provide domain-independent knowledge based retrieval"[4]. They believe that Wikipedia contains knowledge that helps them to match both document and query. This effort of harnessing information from Wikipedia is similar to another project called DBpedia, which will be discussed in Section III. The main functionality for Koru is in its ability to expand user search query based on information obtained from Wikipedia. As a consequence, the search result returned may contain topics that are not stated in the user query. However, Koru’s indexing technique is poor in term of coverage. A search will always tend to return less than 20 results, which show that not enough articles is indexed in Koru. Although the research shows that extra articles that are related to search query could be detected by exploring information harnessed form Wikipedia, the research failed to show that it could be implemented on a large scale of articles. Due to this, it is not known whether the same concept could be applied to the World Wide Web without losing the accuracy and precision of the search result.

Another research based on semantic search is Look4. It is a meta search engine that aims to enhance the result from a web search engine by using a network of concepts that are built from of the WordNet ontology [5]. They claimed that the categorization of the other search engines is done based on statistics and covers mainly the most common cases. In so doing, it decreases the coverage of all possible meanings. To solve this problem, Look4 uses WordNet as the resource. The main idea behind Look4 is to "understand" what kind of information a user is looking for, with the help of ontology and natural language processing. The research claims that Look4 could achieve approximately 60% increased precision level when they make an evaluation compare with Google, Yahoo and Bing for a set of 50 selected keywords.

III. THE STRUCTURE OF THE META SEARCH ENGINE

We propose another alternative approach to building a meta semantic search engine. The term meta is used here to indicate that we make use of Google as the underlying search engine rather than building our own search engine. The main idea of this meta search engine is based on Yippy, which is a meta search engine that has the ability to cluster search result to a group of topics that may be relevant to the user. As mentioned, Google will be used as the search engine and DBpedia is used as its knowledge base. By combining the functionalities of Yippy, Google and DBpedia, user will experience another form of searching that will deliver quality search result and generate related concepts based on the meaning of the query. Figure 2 shows a high level architecture of the proposed meta search engine.

There are four main stages involved in search process for our meta search engine:

1) Receive a search query from the user. The search query is matched with the knowledge base (DBPedia) to extract the relevance concepts related to the query.
2) Expand and reconstruct the search query and resend it to Google.
3) Process the search result from Google. The search result is rearranged and clustered based on relevant concepts detected from the user query.

Figure 1: The QEDEX System
In this work, we aim to achieve several objectives:

1) Build a meta search engine that utilizes a knowledge base to expand the search result provided by a non-semantic search engine with a set of clustering process.

2) Study the potential of DBpedia the knowledge base to enrich the search result.

3) Compare the search result generated with the results obtained from well known search engines.

The proposed semantic meta search engine here is different from the pure semantic search engine in that the indexing of the web pages does not involve any notion of semantic. However, this research is based on the belief that a semantic meta search engine is capable of performing a quality search as long as the non-semantic search engine can provide a quality search result. This is analogous to a scenario where a user tries to search for information using Google. When he wants to find information relating to "cat" he would type in the query with the string "cat". Upon viewing the result provided by Google, if he is not satisfied, he would type in another query such as "cat breed" and another result will be generated by Google. This process is repeated until he finds the right information. Using the same theory, a semantic meta search engine can help the user to search for more information by considering all the related concepts associated with "cat". This work is motivated by the fact that existing semantic search engines are still not able to deliver the desirable result. As an illustration, if a simple query of "what is gif" is typed in Hakia, it will show "Welcome to Gifs.net!" as the top result.

Based on query that was submitted on the 24 Jan 2011.

A. DBpedia

In general, DBpedia is a community that tries to extract structured information from Wikipedia to a form that can be queried just like a database. Wikipedia, being a large source of information available on the World Wide Web, consists of descriptions of vast amount of concepts. This description of concept can be a useful source of knowledge available in web. For example, various efforts are done to enhance information in Wikipedia such as [6], [7]. This knowledge contained in Wikipedia has been extracted by community of DBpedia into a form of RDF triple as defined by W3C. As such, semantic technology can be used to explore and extract any information contained in it. Figure 3 shows example of how data can be extracted from DBpedia [8]. As shown in Figure 3, every page from Wikipedia will have similarity in terms of page structure. For example, a page will start with a title name, followed by a short description on the specific title, then an index table and so on. The bottom part of Wikipedia page shows the categorization of a topic. Furthermore, the label of a topic in various languages is captured on the left hand side of the page. Some Wikipedia pages even provide infobox on the right hand side of its page. Since all these information is structured, it is possible to extract such information for every page in Wikipedia and generate a huge graph based on the concept.

The DBpedia dataset includes an estimated 1.95 million concepts [9] and it is still growing. As of March 2010, the DBpedia dataset describes about 3.4 million of "things" and over 1 billion of "facts". Since the information extracted in DBpedia follows RDF triple standard, one can just retrieve any information provided by DBpedia by just querying the database. The implementation of DBpedia dataset enables one to write SPARQL queries like: "Give me all sitcoms that are set in NYC" or "Give me all tennis players from Moscow" [10]. More details on how DBpedia being applied in the proposed meta search engine will be discussed in Section IV.
B. Google and Yippy

Google is currently the most popular search engine. Although it is able to provide a good quality search result, the matching of the search result focuses on user query string only. For example, one could not get a list of search result focused on “gamma distribution” if the search query is for the term “gamma”. In other word, the quality of the search result will have some dependency on the user knowledge on specific domain. Even though Google does provide some search suggestions at the bottom of the search result, user still needs an extra click to get the additional information. Hence, the meta search engine here will try to enrich search result from Google by applying DBpedia.

Yippy on the other hand is a meta search engine that is able to provide a search result which is clustered in a group of common keyword based on the occurrences of word in the search result. This allows the user to perform a deep search based on his previous search query. However, the clustering in Yippy is not based on the semantic of the query string By clustering a search result based on the meaning of the search query, user should be able to get the relevant information faster and painlessly. The proposed work will capitalize on the limitations of Google and Yippy to produce a better meta search engine.

IV. FIRST IMPLEMENTATION

The current implementation of the semantic meta search engine is called Meow. The web server is developed using Java whereas DBpedia is loaded to AllegroGraph server which is an RDF triple store database. A few datasets are used in Meow to enrich the search result and more datasets are being experimented as well. In the current stage, the datasets that are used in DBpedia include:

1) Article to Category dataset enables Meow to figure out which domain of topic a search query should belong to. For example, given a term such as “artificial intelligence”, Meow will detect related concepts such as “cybernetics” and “formal sciences”.

2) Disambiguation dataset enables Meow to extract the possible meaning of the query. For example, based on DBpedia, the term “gamma” refers up to 10 concepts, like “gamma wave”, “gamma correction” and “gamma distribution”.

3) Labeling dataset enables Meow to understand how a certain concept can be referred to in other languages. Applying labeling dataset is helpful in searching for certain information that is independent of the language such as picture.

4) Page link dataset enables Meow to detect any relevance concept that a term may be related to. For example, the concept of “artificial intelligence” is linked to page such as “knowledge representation”, and “logic programming” in Wikipedia.

5) Redirect dataset enables Meow to extract all the equivalent terms that refer to the same concept, which may be helpful in guessing the meaning of the query. For example, “AI” and “artificial intelligence” refer to the same concept so both terms are bounded under the dataset of redirect.

The snapshot of the implementation of Meow is shown in Figure 4 and Figure 5. Figure 4 shows the general search interface of Meow. The interface is divided into three parts. The left hand side contains suggestion that a user may be interested in based on the relationship obtained from DBpedia. The middle section contains the general search result extracted from Google. This result has been re-ranked by considering the occurrence of related word from the original search query. For example, “gamma” is related to “physics” in some sense, and so the term “physics” has influence on the result ranking. On the right hand side of the page, the search result is clustered based on the related concepts obtained from DBpedia. This allows user to explore the search result in a more specific domain.
even deeper search based on a specific topic that he may be interested in. On the other hand, the right hand side of the page shows topic that are clustered by Meow based on concepts that are related to "gamma" in DBpedia. In the current implementation, most of the concept considered in clustering is based on page link dataset since it provides linkage between concepts in different domain of the topic.

![Table of Gamma-related concepts]

**Figure 6.** A close view on suggestion from Meow

### V. CONCLUSION/FUTURE WORK

In this paper, we presented Meow, a semantic meta search engine that is able to transform a query from a non-semantic search result into a semantic search result. This is made possible through the exploration of concepts in DBpedia. DBpedia is used as a knowledge base because it contains a large dataset that is able to describe general concepts for various domains. By applying DBpedia, Meow is able to enrich the search result from Google. However, the information that can be queried from DBpedia for certain topic may be large and may include concepts that are not relevant. For example, artificial intelligence is both linked to concept of "MIT" and "knowledge representation". However, DBpedia has no way of indicating which concept is more relevant. At the moment, Meow considers all concepts as relevant as long as a relationship is detected. For future work, we aim to investigate various strategies to utilize DBpedia dataset to improve this problem. We will also conduct experimental evaluations to compare the performance of Meow against other search engines.

### ACKNOWLEDGEMENT

The first author is supported by MOHE, FRGS grant.

### REFERENCES


