Retrieving Information about Real World Activities from the Web

Karapetsas Eleftherios[†] Yusuke Fukazawa[†,††]
Jun Ota[†]

The web is thriving with websites containing How-To articles and DIY guides. In websites like that information about activities and guides how to accomplish them are located. An activity is defined as any action that a person can accomplish. Our research focuses on Retrieving activities from the Web related to a user's search query. To accomplish that we have created a system that performs meta-search in multiple How-To websites using semantic query expansion. An overview of the activities retrieval system is presented along with the explanation of the query expansion algorithm which utilizes ConceptNet. Finally experimental results gathered from a user study are given in order to evaluate the performance of our system.

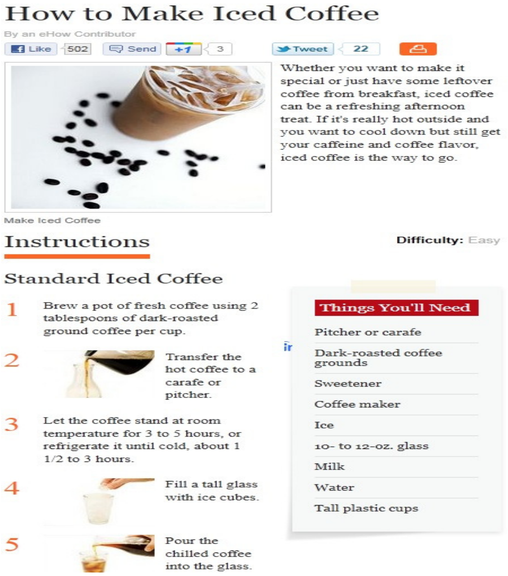
1. Introduction

The World Wide Web is a place where a huge amount of information is being circulated. A very big part of this information is concerning activities that a person can do. These activities can mainly be categorized as:

- Actions that someone can accomplish
- Tasks to be completed
- Things that can be created

Activities like the above are mainly located in websites providing know-how knowledge, the so called How-To sites. The web is thriving with websites containing know-how information and how-to articles. These articles span a great range of topics from home decoration to mechanical engineering.

Such websites are either populated manually by a closed group of reviewers or are open for content submission by anyone in the world. Each website has some categories that it is more focused on but all of them tend to cover a wide range of activities. In Figure 1 a screenshot of popular site eHow.com can be seen providing a typical article with instructions on how to accomplish an activity. As can be seen the steps that have to be taken are numbered making it easier to understand and follow the guide. Ehow in particular provides a certain difficulty score in each article explaining how hard or easy the guide is to follow and also a number of items that would be needed in order to successfully accomplish the activity detailed inside the guide.



How to Make Iced Coffee

By an eHow Contributor

Whether you want to make it special or just have some leftover coffee from breakfast, iced coffee can be a refreshing afternoon treat. If it's really hot outside and you want to cool down but still get your caffeine and coffee flavor, iced coffee is the way to go.

Make Iced Coffee

Instructions

Difficulty: Easy

Standard Iced Coffee

1 Brew a pot of fresh coffee using 2 tablespoons of dark-roasted ground coffee per cup.

2 Transfer the hot coffee to a carafe or pitcher.

3 Let the coffee stand at room temperature for 3 to 5 hours, or refrigerate it until cold, about 1 1/2 to 3 hours.

4 Fill a tall glass with ice cubes.

5 Pour the chilled coffee into the glass.

Things You'll Need

- Pitcher or carafe
- Dark-roasted coffee grounds
- Sweetener
- Coffee maker
- Ice
- 10- to 12-oz. glass
- Milk
- Water
- Tall plastic cups

Figure 1 A typical How-To article on ehow.com

In these websites people share their experiences by providing How-To articles or DIY guides. (DIY stands for Do It Yourself) Websites like that are extremely popular and have been getting more popular with the passage of time. In Table 1 we present a small overview of various How-To sites and their characteristics. The statistics come from Alexa the most prestigious web statistics and Information Company.

[†] The University of Tokyo
[††] NTT DOCOMO, Inc.

Table 1 Some How-To Websites and their Characteristics as of August 2011

Website	Number of Articles	Average Number of Users Per Month	Alexa Worldwide Traffic Rank	Range of topics covered
ehow.com	2.000.000 +	800k-5million	128	General How-To for wide variety of topics
howtodothings.com	52.500	80k-300k	12.142	General How-To, Business, Health, Travel
instructables.com	2.000.000+	200k-1million	1.243	DIY, creating anything with you hands
hackaday.com	28.300	80k-300k	10.814	DIY, Software, Hardware Guides
makezine.com	50.600	200k-1million	7.474	DIY, mechanics,engineering
about.com	5.880.000	5million+	69	How To, Advice, Information
howto.wired.com	10.900	800.000	601	Gadgets, Hardware, DIY

As you can see from Table 1 there is a big number of How-To sites but there is not a concrete way to search all of them such as a how-to meta-search engine. One would need to search each site individually which would be a very tedious and manual process because no Meta-search engine for How-To sites exists.

In our research we set out to gather activities from how-to sites like the above related to a search query that a user performs. Our goal is to provide the users with activities related to their query using the abundance of information located in said sites. Retrieving activities from the web is important because:

- People often have the need to find instructions on how to accomplish an activity, and using the web is the fastest and easiest way to achieve that.
- How-To sources are peer-reviewed and edited by many people, so reviewed information from people all over the world can be acquired.
- Combining information from multiple websites can cover every aspect of an activity and provide a wealth of information that cannot be found from other sources.

For us activities are any real world task that consists of actions and objects. An action is represented by a verb and an object is represented by a noun. So example activities could be “Make Ice Coffee”, “Paint the wall” with the verbs and nouns clearly stated. Activities like that can easily be extracted out of How-To and DIY websites because the very nature of those websites is to provide their users with detailed information on how to accomplish challenging undertakings and how to create interesting things.

2. Related Research

There are many instances in the literature where people attempted to retrieve activities from the web. In [1] they created a system of activities and services that helps users to solve real life problems based on their location and time of day. Their system retrieves possible activities to satisfy the user’s needs for some services. In [2] they made a system that retrieves possible activities that a user can perform depending on the place or on the domain of the task. These two approaches have the problem that they manually mine the activities to populate their knowledge bases, which is not cost-effective and not easy to scale up in a bigger application.

In other approaches, such as [3] they used the results from multiple websites by creating a meta-crawler of multiple search engines. Finally in [4][5] they used query expansion by semantic analysis using DBPedia, the machine-understandable Wikipedia. The shortcoming of these approaches is that in [3] they directly matched the user’s query giving few results, while in [4][5] they did query expansion but using DBPedia the expansion was too wide and not activity-based.

In [8] they used query expansion with the same tool as we will present in the next Section, ConceptNet, but their work was focused on finding related keywords in a general context and not with a specific purpose such as activities is for us. They combined Wordnet with ConceptNet for query expansion. Wordnet achieves expansion by finding synonyms, the so called synsets. On the other hand in [9] they used ConceptNet for query expansion but applied it in the context of semantic query expansion for finding related images and without giving any details on their query expansion algorithm.

3. Our Approach

3.1 Requirements

We propose the creation of a system which retrieves activities from multiple How-To websites, related to a query that the user will input. In order to thoroughly cover the domain of the user's search we are using query expansion based on the part of speech of the user's query. The requirements of our system are to provide the user with information about and links to activities that can be performed regarding their search query.

It has been observed that many times it is not easy for users to describe activities because of insufficient knowledge that a user might have about the activity domain or simply because the user is not certain of the exact nature of activities he would like to learn about.

Since the user is expected to not have a good understanding of the search domain an additional requirement of our system is to provide us with as many relevant resulting activities as possible in order to satisfy the user's need for knowledge about the query. We believe that this is accomplished by the query expansion algorithm.

3.2 An example scenario

Suppose that Anna is holding a party for her birthday and that there are only 4 hours left until the party starts. With limited time in her hands Anna wants to impress her guests and knows that all of them are big fans of good wine. So she decides to impress them by doing something for them related with wine. Using our system and with the search query "wine" she will retrieve activities such as "Turn a wine bottle into a lamp" "Make wine Jelly" "Create your own wine". If she tries to perform any of these activities her guest will surely be impressed.

3.3 Description of our system

In section 2 we noted that related research had the shortcoming that they manually populated their knowledge base which was both not cost-effective and also time consuming. In our system we use information located in multiple How-To websites so we are not using a single knowledge base but by extracting information from the web we utilize knowledge from people all over the world. This provides us with a much broader amount of activities, satisfying the requirement set in 3.1.

Differentiating further from other related research as noted in Section 2 our approach does not directly match the user's query. Instead we are using common sense knowledge obtained from ConceptNet [7]. ConceptNet is a knowledge database created by MIT which stores common sense knowledge in the form of relations between words and concepts. An example of relations that are used is UsedFor(wine,drinking) LocatedIn(kitchen,house). The semantic knowledge base of ConceptNet is populated manually by people and practically anyone has the ability to vote up relations that are accurate and vote down relations that don't make sense or are ambiguous. Users of the knowledge base use assertions to see if a relation exists or not. All assertions have scores associated with them which dictate how probable they are of happening in the real world.

In contrast with direct matching our approach with ConceptNet provides a lot more results quite relevant to the original query further satisfying our requirement for a big number of results. Furthermore unlike [8] we do not use WordNet in anything other than syntactic analysis because for the purpose of finding activities, synsets do not provide a good enough expansion. What is needed is semantic expansion, which we achieve through ConceptNet. Our approach consists of the following stages which can be also seen in Figure 2.

1) *Syntactic analysis:* Is performed on the query to extract important syntactical data such as what part of speech each word is. Wordnet lexical database is used for this purpose [6].
2) *Query expansion:* We expand the query by finding concepts related in various ways to the original. This is accomplished by using ConceptNet. ConceptNet allows people to search for common sense information regarding some words in the form of assertions.
3) *Crawling of How-To websites*: Various How-To websites are then queried for both the original term and the related concepts. This stage returns activities related with the original query in interesting ways.
4) *Filtering and Ranking of Results:* Finally results are filtered to eliminate duplicates, ranked according to relevance and they are presented to the user inside a GUI program created by us.

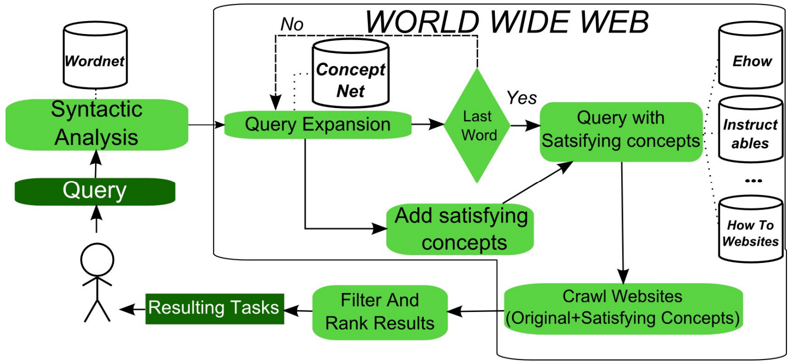


Figure 2 Overview of our system

During the syntactic analysis the query is separated into its constituent words. All words part of speech is tagged, prepositions are removed and plurals are turned into singular. This is done in order to prepare the query for query expansion.

For the query expansion stage we also use ConceptNet just like in [8] but unlike their research we do not use the get_context() tool provided by ConceptNet since we are not interested in general related concepts to the original user's query. Instead we are performing only specific assertions depending on the identified part of speech of the word in order to get related concepts that would help with retrieving activities.

So depending on what parts of speech were identified from syntactic analysis, ConceptNet is queried with various questions in order to obtain common sense information as can be seen in Figure 3. If the query is a:

- Noun: What is the noun used for? What possesses this noun? What is the noun capable of? What would you find in the noun? Where would you find the noun?
- Verb: What can receive the verb as an action? What is capable of doing this verb? What does the verb cause? What can the verb create?
- Adjective/Adverb: If ConceptNet contains information about the combination of adjective-noun or adverb-verb then it is extracted. If not the adjective and adverb does not undergo query expansion.

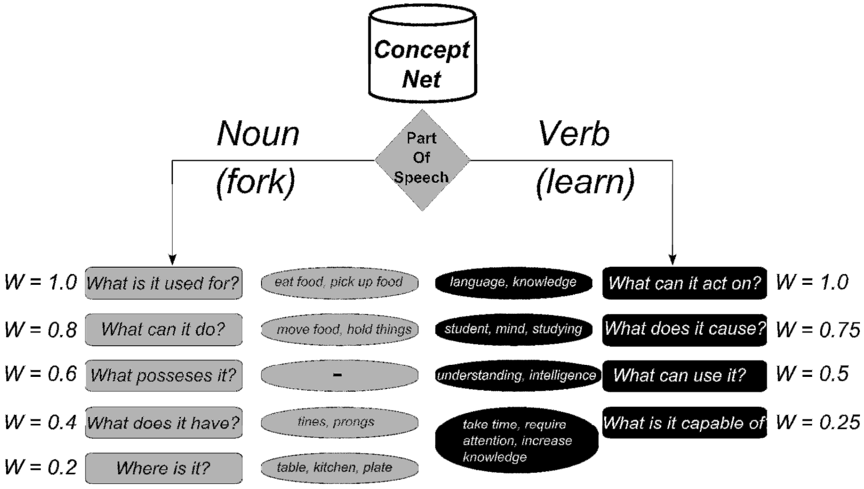


Figure 3 Query Expansion with ConceptNet

All this returns related concepts which are given a relevance score. The algorithm used for query expansion scoring is a modified form of Spreading Activation[10] used almost universally in Query Expansion literature. In our approach we created a smaller semantic tree for query expansion and the weights of each relation are already set beforehand instead of decaying by distance such as is usually the case. The reason for this is that we focus the expansion on terms and relations that are more likely to produce activities related keywords.

The weights of each relation can be seen in Figure 3 next to the relation. These values were decided after using various values in tests of our system and observing the returned results. The primary metric that guided us towards these values was the number of returned activities which contained verbs when queried with nouns and nouns when queried with verbs. Secondary metric, used in the case that the primary metrics of two queries were equal, was the actual span of concepts that the activities covered. Naturally as can be observed by the values they push the related concepts into forming a Noun-Verb combination which is what activities are composed of. Further in the future these values can be optimized through a user evaluation experiment. In Figure 4 examples of the calculations of these values by the expansion of a query with the word "Map" can be observed.

$$\sum_{i=0}^{i*j} conceptScore(i,j) = s(i) * w(j)$$

Equation 1 Query Expansion algorithm

Using Equation 1 the calculation of the relevance *conceptScore* of each concept is accomplished. We have a concept *i* returned by assertion relation *j*. *s(i)* is the score given by ConceptNet to the specific concept returned by the assertion and *w(j)* is the weight value we give to the relation itself.

To present a few examples of calculating the concept scores let's consider the following queries: map, coffee, book and party. By querying ConceptNet for "What is map used for?", which has weight value 1.0, we get "to navigate" with score 4 so the final *conceptScore* is 4. By querying "What can coffee do?", which has weight value of 0.8, we get "stain clothes" with score 2 so the final *conceptScore* is 1.6. By querying "Where can you find a book?", which has weight value of 0.2, you get "shelf" with score 16 so the final conceptScore is 3.2. From the last example one of the reasons why smaller weights for location relations are needed can be determined. Spatial relations tend to have bigger scores in ConceptNet than action relations so a smaller weight is one way to push query expansion towards actions related keywords. Finally by querying "What does a party have?", which has a weight value of 0.4, you get "booze" with score 5, which gives a *conceptScore* of 2.

Out of all the scored concepts we have gathered from expansion we choose the N top, where N is a number determined by us. Again in Figure 4 a part of the semantic tree created by query expansion of the noun "Map" can be seen presenting the related concepts along with their computed *conceptScore*.

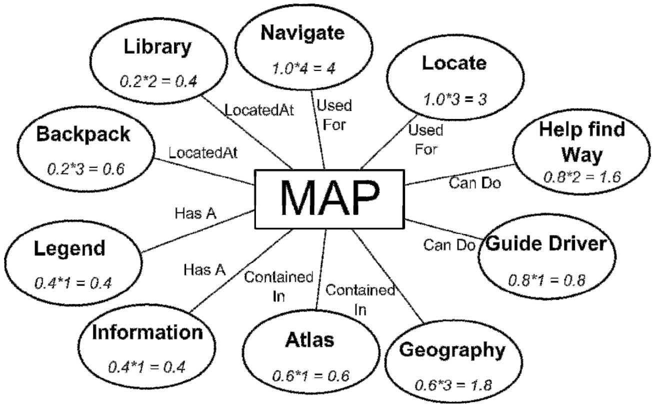
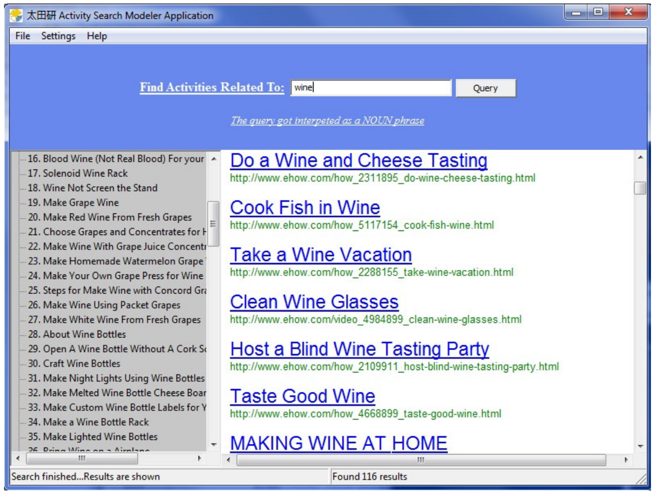


Figure 4 Semantic Tree generated by query expansion for “Map”

Subsequently we proceed to the final stage which is to actually query the websites with the chosen concepts from query expansion. The websites are queried with both the original query and its combination with the related keywords that have a high enough *conceptScore*.

The results are filtered and all duplicate activities are removed. Then they are ranked depending on the *conceptScore* of the concepts that returned each activity. Furthermore since each website has its own ranking system we use that as an additional ranking metric for the resulting activities if we have equal concept scores.

Finally the activities are presented to the user through a GUI interface that we have created as can be seen in Figure 5. In the top part of the GUI is the query textbox where the user can enter his query. On the bottom left column the activities list is presented while on the bottom right the links to the activities articles along with their urls can be seen. Under the query's textbox the user can see what the main syntactic interpretation of his query is.



Find Activities Related To: wine

The query got interpeted as a NOUN phrase

16. Blood Wine (Not Real Blood) For your
17. Solenoid Wine Rack
18. Wine Not Screen the Stand
19. Make Grape Wine
20. Make Red Wine From Fresh Grapes
21. Choose Grapes and Concentrates for H
22. Make Wine With Grape Juice Concentr
23. Make Homemade Watermelon Grape
24. Make Your Own Grape Press for Wine
25. Steps for Make Wine with Concord Gra
26. Make Wine Using Packet Grapes
27. Make White Wine From Fresh Grapes
28. About Wine Bottles
29. Open A Wine Bottle Without A Cork Sc
30. Craft Wine Bottles
31. Make Night Lights Using Wine Bottles
32. Make Melted Wine Bottle Cheese Boar
33. Make Custom Wine Bottle Labels for Y
34. Make a Wine Bottle Rack
35. Make Lighted Wine Bottles

Do a Wine and Cheese Tasting
http://www.ehow.com/how_2311895_do-wine-cheese-tasting.html

Cook Fish in Wine
http://www.ehow.com/how_5117154_cook-fish-wine.html

Take a Wine Vacation
http://www.ehow.com/how_2288155_take-wine-vacation.html

Clean Wine Glasses
http://www.ehow.com/video_4984899_clean-wine-glasses.html

Host a Blind Wine Tasting Party
http://www.ehow.com/how_2109911_host-blind-wine-tasting-party.html

Taste Good Wine
http://www.ehow.com/how_4668899_taste-good-wine.html

MAKING WINE AT HOME

Search finished...Results are shown

Found 116 results

Figure 5 the GUI of our system

4. Evaluation

In order to evaluate the performance of our system we conducted a small study by asking 20 people to use our application. They were asked to perform 5 queries using (i)Google (ii)Our System (iii)Our system without using ConceptNet. The 3rd option was added to evaluate our choice of ConceptNet for query expansion. They were asked how many interesting activities results they had from each search method out of the first 40 results given by each. The subjects were encouraged to use nouns and adjective/nouns combinations since at this first stage of evaluation of our system we mainly wanted to test the noun query expansion algorithm.

Our test subjects had various different backgrounds in both science and literature and came from many different countries all over the world. All subjects are residents of an international dormitory located in Tokyo. By the results of the study we both verified the strengths of our system but also its shortcomings.

In Figure 6 the interesting results averages from all 20 people for each method are displayed along with the corresponding standard error. From that figure it is obvious that for finding interesting activities when querying a noun word, our system has an advantage over traditional search engines such as Google since the interesting results average is almost double that of Google. Moreover ConceptNet provided a 6.9% increase in interesting activities thus justifying the need for query expansion to solve the activities retrieval problem.

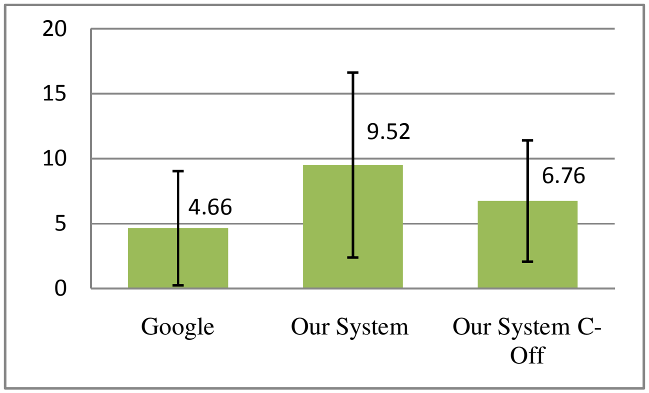


Figure 6 Interesting results averages of the user study

In Table 2 we provide some example queries made by users during the evaluation and the number of interesting results each method returned. Our users considered as interesting activity anything that would motivate them to get off the computer and try to perform it as soon as they read the article.

The first 4 queries are examples of good queries that are suitable for our system. When a simple one word query is given our system explores the domain of the word and returns relevant results focused on activities. The activity that users considered the most interesting for every query is also shown. Most users stated that they were positively surprised by the resulting activities and that they did not expect to get such interesting and varied results.

The last 2 are example of queries not suitable for our system. Advanced concepts and very specific words such as "gesture recognition" are not suitable for query expansion by ConceptNet since it is not considered common sense knowledge. Moreover when the user knows exactly the activity he wants to perform, standard search engine search seems to be a lot more accurate since it will return more results as can be seen by "calibrate monitor" query.

Table 2 Representative uses of our system

Query	Google	Our System	Our System ConceptNet-Off	Interesting Activity Example
pencil	6	31	14	Make a pencil crossbow
coffee	2	15	12	Dye hair with coffee
camera	4	8	6	Camouflage a camera
eggplant	3	7	7	Make grilled eggplant sandwich
gesture recognition	4	0	0	Perform hand gesture recognition
calibrate monitor	29	6	6	Calibrate your display

5. Conclusions

We have presented a system to mine the web for interesting activities by utilizing query expansion using ConceptNet and multiple How-To websites. As a result of our experimental study we can conclude that the presented method is almost twice as effective as simple Google search as far as exploring and discovering interesting activities is concerned.

Remaining work for the future is to improve the algorithm that ranks the final results and conduct further evaluation on both the verbs and verb-noun queries. Moreover user evaluation to determine better optimized relation weight values during ConceptNet assertions would be beneficial. Furthermore we would like to compare the results of our system against simple search of the How-To websites and evaluate further how many interesting resulting activities our system would return compared to normal How-To website search.

Finally we have set as our goal to proceed with machine learning and interpret the resulting activities into a machine understandable form. This way the system can then be employed in robotics without the need for a human to manually input expert knowledge into a robot. Instead using our system a robot can potentially search the internet by itself in order to learn how to accomplish activities and then act on that acquired knowledge.

References

1) Naganuma et.al. “Task Knowledge Based Retrieval for Service Relevant to Mobile User’s Activity” In Proc. ISWC2005, pp. 959–973, 2005.

2) Fukazawa et.al. “Proposal and User Evaluation of Enhanced Task-based Mobile Service Navigation System,” Transactions of Information Processing Society of Japan,Vol.50, No.1, pp.159-170, 2009.

3) Shekhar et.al. “An Architectural Framework of a Crawler for Retrieving Highly Relevant Web Documents by Filtering Replicated Web Collections”, International Conference on Advances in Computer Engineering, pp.29-33, 2010.

4) Mirizzi et.al. “Semantic wonder cloud:exploratory search in DBpedia”, in Proc. ICWE’10, pp.138-149, 2010.

5) Mirizzi et.al. “From exploratory search to web search and back”, in Proc. PIKM’10, pp.39-46, 2010.

6) Christiane Fellbaum.”WordNet: An Electronic Lexical Database.” Bradford Books 1998.

7) H. Liu et.al. “ConceptNet – A Practical Commonsense Reasoning Tool-Kit”, BT Technology Journal/ Vol. 22 Issue 4, pp.211-226, 2004.

8) Ming-Hung Hsu, Ming-Feng Tsai and Hsin-Hsi Chen, “Combining WordNet and ConceptNet for Automatic Query Expansion: A Learning Approach” , 4th Asia Infomation Retrieval Symposium, AIRS Vol. 4993 pp.213-224, 2008

9) Roohullah, K. and Jaafar, J, “*Semantic Query Expansion Using Knowledge Based for Images Search and Retrieval.”* International Journal of Computer Science & Emerging Technologies Vol. 2 pp. 1-5 , 2011

10) Salton, G. and Buckley, C. “On the Use of Spreading Activation Methods in Automatic Information Retrieval” In: Proceedings of the 11th CM-SIGIR Conference on Research and Development in Information Retrieval. pp. 147-17 , 1988