# TEXT DOCUMENT INFORMATION RETRIEVAL BASED ON CONCEPTS

**Amir Hamzah**
Institute of Science and Technology AKPRIND,
Jalan Kalisahak 28, Yogyakarta 55222, Indonesia
Phone: (0274)-563029; Fax: (0274)-563847
e-mail: miramzah@yahoo.co.id

**ABSTRACT**

The huge volume of digital information collected automatically by internet technology has caused problems in information retrieval. Finding the right information from a large collection is very difficult. The difficulty in most search engines are caused by a string matching algorithm that return a match whenever an exact occurrence of the search term is found. To address this problem and considering that the document collection is not only a collection of words but also a collection of concepts, we promote a new technique of information retrieval that is based on concepts.

The difference between word-based and concept-based technique are indexing and retrieval. During indexing, this technique classifies documents into concepts extracted from the collection via clustering technique to construct concept indexing besides term indexing. During retrieval, this techniques ranks document base on a combination of term and conceptual similarity, in the formulation of *doc-score = $\beta$ * conceptScore + (1-$\beta$)*TermScore* where $\beta$ is the weight of concept score. The clustering algorithm is chosen from partitional model that linear in complexity, that is Bisecting K-Means.

Two kinds of test collections, i.e. text document of news (1000 and 3000 news documents), and text document of academic articles (1000 academic abstract in information technology) were used to conduct the experiment. Performance evaluation was measured using average precision and R-precision.

The results of the research showed that by setting $\beta$ =0.5 to $\beta$ =0.9 would improve significantly the precision of concept-based approach over the word-based only ($\beta$ =0). The improvements are about 5.2% to 8,3% for average precision and 16.9% to 31.5% for R-precision.

***Keywords**: information retrieval, concept-based, word-based, precision*

**INTISARI**

*Melimpahnya informasi digital yang dikoleksi secara otomatis oleh internet telah menimbulkan problem dalam temu kembali informasi. Menemukan informasi yang tepat dalam koleksi dokumen yang besar adalah sangat sulit. Kesulitan ini disebabkan karena pada kebanyakan mesin pencari berbasis pada pencocokan string sehingga akan otomatis memberikan dokumen sebagai jawaban jika terdapat string yang cocok. Untuk menangani hal ini dan dengan mengingat bahwa dokumen bukan saja merupakan koleksi kata tetapi juga merupakan koleksi konsep, penulis mengusulkan teknik baru temu kembali informasi yang berbasis pada konsep.*

*Teknik ini berbeda dengan temu kembali berbasis kata pada tahap indexing dan tahap temu kembali. Pada tahap indexing teknik ini mengklasifikasi dokumen berdasarkan konsep menggunakan teknik clustering untuk menyususn index konsep disamping index kata. Pada tahap temu kembali, teknik ini merangking dokumen berdasarkan kombinasi similaritas kata dan konsep, dalam suatu formula doc-score = $\beta$ * conceptScore + (1-$\beta$)*TermScore dimana $\beta$ adalah bobot skor konsep. Algoritma clustering dipilih dari model partisi dengan kompleksitas linear, yaitu model Bisecting K-Means.*

*Percobaan dilakukan pada 2 jenis koleksi, yaitu dokumen teks berita (1000 dan 3000 dokumen) dan dokumen akademik (1000 dokumen abstrak bidang IT). Evaluasi kenerja temu kembali diukur dengan rata-rata presisi temu kembali dan R-presisi.*

*Hasil penelitian menunjukkan bahwa dengan men-set $\beta$ =0.5 to $\beta$ =0.9 menunjukkan peningkatan precisi berbasis konsep terhadap basis kata ($\beta$ =0). Peningkatan precisi adalah sebesar 5,2%sampai 8,3% untuk rata-rata precisi dan 16.9% sampai 31.5% untuk parameter R-precision.*

***Kata kunci**: temu kembali informasi, basis konsep, basis kata, presisi*

## INTRODUCTION

The recent years progress of computer technology has caused an explosion of electronic information published on the internet. The volume of the information was growing rapidly each year. In 2011 Google collected more than 25 billion pages (http://www.google.com, 2011). However, the huge amount of such information, makes it practically impossible for human user to be aware of much of it. This problem of abundance information has triggered another problem in information retrieval, either in searching the relevant information or in presenting the search results.

Mostly information retrieval tools use keyword search, which is unsatisfactory option because of its low precision and recall. Problem that may arise from matching technique is related to searching the relevant documents. The effect of polysemy of words, where a word may have multiple meaning, has caused many unrelevant documents to be retrieved. On the other hand, word synonymy have caused many relevant documents in the collection can not be retrieved because they use different words with the query's word.

Many methods have been proposed to solve the problem of low precision search engine. One of the approaches is by considering the semantic aspect of documents. This approach views documents as collection of concepts and therefore matching query and document must be done in the level of concept.

### Related Work

Concept-based approaches to information retrieval has become state of the art in search engine researchs. Even this approach is also applied in other disiplin other than IR such as software comprehension (Cleary et.al.,2008). Although there are differences in that how to construct the conceptual structure in the implementations there is an agreement that the conceptual approach is better than word-based approach. Some systems use conceptual taxonomy as conceptual structure such as Sun Microsystems Conceptual Indexing (Woods, 1997) or using explicit semantic analysis that combine word-based and concept-based (Egozi 2010; Egozi et.al,2011). Many systems use domain ontology such as Ontoseek (Guarino et.al., 1998), Ontobroker, Readware Consearch (Haav et.al.,2005) and applied concept-based Ontology in Cross-Language Search (Rad et.al.,2010). Other use thesaurus and linguistic network such as Thunderstone MataMorph and Thunderstone Webinator or conceptual network (Zeng et.al.,2003; Widyantoro,2007). Goyal et.al. (2009) implemented fuzzy logic to automatically construct concept from concept hierarchies.

The Among other systems, using domain ontology as conceptual structure is the model that much more likely to be chosen because of its advantages in implementations (Gruber,1995; Van Height et.al.,1997). However ontology also has disadvantages such as pre-existing dictionary often do not meet the user's need for interesting concept (Guarino et.al.,1998). Using ontology also need human effort to construct it manually, that very time consuming and intensive labor.

In this study, another approach that we choose is extracted concepts from document using clustering technique. It has been successfully done by Karypis and han (2000) through concept indexing where the concept is defined as cluster center of clustering result.

### The Concept of Information Retrieval

Formally IRS is defined as a system to store, to manage and to retrieve information. As the forms of information becomes spread to many forms such as text, images, audio and video the IR has become more challenging research topic. However, as text is the dominant form of information, a Text Information Retrieval System (TIRS) still need improvement to force many problems such as the explosion of text available in the web.

According to the way of matching documents representation and query representation, two models have been developed, i.e. word-based information retrieval model and concept-based information retrieval model (Haav et.al.,2005). The main principals of those models are described below.

### A. Word-based Information Retrieval Model

Firstly, IR model commonly used in commercial search engine is based on keyword indexing system., although nowadays concept-based has implemented concept-based in feedback (Jalali and Borujerdi, 2010). Figure 1. shows the scheme of word-based Information Retrieval Model. In this model, keyword list are used to describe content of documents, although it does not say anything about semantic relationship

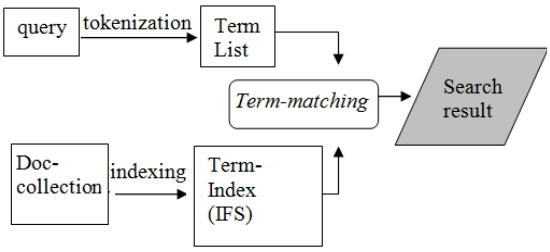


Figure 1. The Scheme of Word-based Information Retrieval System

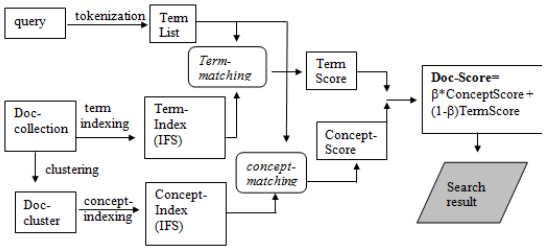


Figure 2. The scheme of concept-based information retrieval system

between words. This is the weakness of this model, because one could easily choose a valid synonymous word that is not in any textual and therefore fail the query search.

By using Inverted File Structure (IFS) to implement the internal data structure, this model can be constructed efficiently. However, it has principal problem in that it does not take into account meaning of the word or phrase. Therefore word in this model is only a sequence of binary code representing a word.

**B. Concept-based Information Retrieval Model**

I The meaning of a text or a word depends on conceptual relationships rather than to linguistic relationship found in dictionary. A new generation IR model is constructed from this idea, where a set of words will be mapped to the concept such that a content of document collection is described by a set of concepts. Concepts can be extracted from the text by categorization. Haav et.al. (2005) and also Egozi (2010) said that the existence of conceptual structure is crucial in this model. Nowadays there are many model to concsttruct the concept from document collections. At least there are five model of conceptual structure in Concept-based IR products that available recently, i.e. *conceptual taxonomy*, *formal or domain ontology*, *semantic linguistic network of concept*, *thesaurus* and *predictive model*. Almost all model use manually construction that disadvantages because its time consuming (Haav et.al.,2005).

*Concept Extraction and Indexing*

Concept extraction from document can be done using clustering techniques. After clustering process, the cluster center vector that is computed from the average vector is vector that represented the documents cluster. This vector can represent the cluster about its content, topic or concept, where its values of elements represent term weight to this concept. Finally each document can be index not only by term-indexing but also by concept-indexing.

*Concept Matching of The Query*

In this Concept-based IR model, query-document matching is based on term-matching and concept matching as depicted in Figure 2. The document-score (ds) as rank to the query is combination of term-score and concept-score with weighted by $\beta$ :

$$ds = \beta^* \text{ conceptScore} + (1-\beta)^* \text{TermScore} \quad (1)$$

Term-score of document D represented by vector of dimension-t to the Query Q is computed using cosine similarity formula :

$$\textit{Cosine}\text{-sim}(Q,D) = \frac{\sum_{i=1}^{t} Q_i D_i}{\sqrt{\sum_{i=1}^{t} (Q_i)^2 \sum_{i=1}^{t} (D_i)^2}} \quad (2)$$

Concept-score of document D to the Query Q is computed almost in the same style, that is :

$$\textit{Cosine}\text{-sim}(Q_C,D_C) = \frac{\sum_{i=1}^{C} Q_{Ci} D_{Ci}}{\sqrt{\sum_{i=1}^{C} (Q_{Ci})^2 \sum_{i=1}^{C} (D_{Ci})^2}} \quad (3)$$

Where $Q_C$ is query-concept vector, that is concept-vector that best-match to the query and $D_C$ is document-concept-vector, that is vector that its each elements represents concept weight to the document. Number of concept C, is the value that had been determined before indexing was constructed. For partitional algorithm this value is set by assumption or prediction about how many concepts should make sense to be available in the collection.

**Experimental Setting**

In this study we use three test-collections, two collections are from news in Bahasa Indonesia : N1000 and N3000 and one collection is from academic environment, i.e. abstract collection form field of Information

technology. The details statistic of test-collection is presented in Table 1.

Table 1
Document Collection for test

| Collec-tion | Num of Doc | Num of Word | Num of Index-Word | Num of Cluster | Num of Query |
|---|---|---|---|---|---|
| N1000 | 1000 | 22.543 | 3.092 | 15 | 11 |
| N3000 | 3000 | 18.255 | 3.844 | 21 | 12 |
| Abstrct | 1000 | 8.110 | 2.219 | 17 | 10 |

For each collection we have construct query-list and query relevant judgement as listed in Table 2, Table 3 and Table 4.

Table 2
Query List for News Collection N1000 and N3000

| No | Query | Num of relev doc | |
|---|---|---|---|
| | | N 1000 | N 3000 |
| 1 | Pemberangkatan jamaah haji | 43 | 45 |
| 2 | Pertandingan Piala dunia | 95 | 183 |
| 3 | Pasar uang dolar | 60 | 68 |
| 4 | Penumpasan Gam aceh | 45 | 65 |
| 5 | Kerusuhan ambon maluku | 59 | 75 |
| 6 | Pertandingan tinju Tyson lewis | 25 | 29 |
| 7 | Tki Indonesia di Malaysia | 27 | 35 |
| 8 | Penyelesaian kasus tommy Suharto | 56 | 77 |
| 9 | Pertandingan tenis junior | 31 | 40 |
| 10 | Penyelesian kasus bulog akbar tanjung | 29 | 88 |
| 11 | Konversi minyak tanah | 28 | 90 |
| 12 | Pemberantasan terorisme | | 75 |

Table 3
Query List for Abstract Collection

| No | Query | Num of relev doc |
|---|---|---|
| 1 | Aplikasi logika fuzy | 26 |
| 2 | Sistem informasi | 45 |
| 3 | Jaringan syaraf tiruan | 16 |
| 4 | Pengolahan citra | 12 |
| 5 | Algoritma genetika | 27 |
| 6 | Database | 34 |
| 7 | Sistem pendukung keputusan | 21 |
| 8 | GPS GPRS komunikasi data | 28 |
| 9 | Rekayasa perangkat lunak | 28 |
| 10 | Keamanan system informasi | 14 |

The procedure of tests are as follows :
For each test-collection we build two kinds of indexing, i.e. term-indexing using IFS-structure for word-based IRS and concept-indexing using clustering method for concept-extraction. Clustering method that we choose is partitional algorithm of linear complexity with performance comparably with hierarchical algorithm, that is Bisecting K-Means Clustering (Steinbach, 2000).

Analysis of the IR performance for each model was done by comparison of average Precision and R-Precision between concept-based model and word-based model. The statistical test was done by SPSS for Windows.

## RESULT AND DISCUSSION

The preprocessing steps for N1000 and N3000 test-collection was set by minimum document frequency contain term to be 5. It result in number of term of 3.195 for N1000 and 6.258 for N3000.

*Result for N1000 Test-Collection*

The comparisons between average precision and R-Precision for N1000 test-collection are shown in Figure 3. and Figure4. It can be seen from those figures that the average precision of Concept-based technique are better for almost all query, either for average Precision or R-Precision. The method of Concept-based was set in the weight of $\beta$=0.5. The similarity between query and document is formulated as doc-score = $\beta$* conceptScore + (1-$\beta$)*TermScore. So by setting the $\beta$ to zero would set the similarity purely to word-based technique.

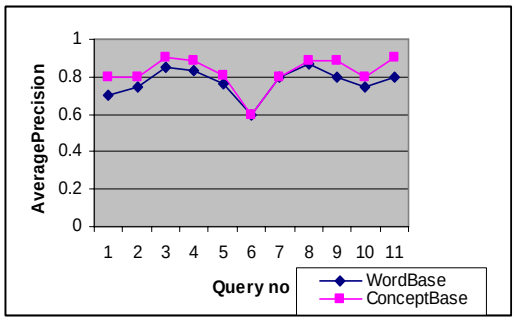


Figure 3. The Average Precision of 11 Query in N1000 test-Collection
(Concept-based Beta=0.5)

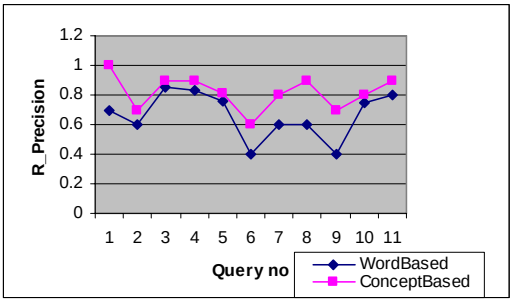


Figure 4. The R- Precision of 11 Query in N1000 test-Collection (Concept-based Beta=0.5 )

The comparison between Concept-based and Word-based for average precision of all queries is shown in Figure 5. In this figure Concept-based is represented by Beta=0.5 and Word-based is represented by Beta=0. The improvement of Concept-based precision is 5.2% for average-Precision and 16,9% for R-Precision.

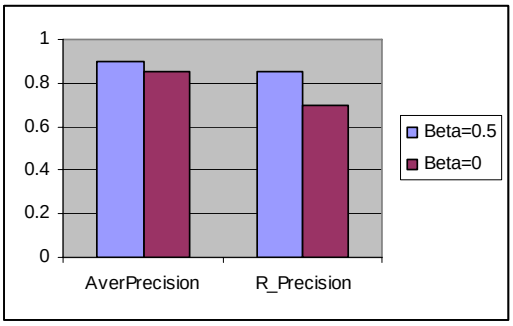


Figure 4. The Avg- Precision and R-Precision of 1All Query in N1000 Beta=0.5 (Concept-based) , Beta=0 (Word-based)

*Result for N3000 Test-Collection*

The comparisons between average precision and R-Precision for N3000 test-collection are shown in Figre 5. and Figure 6. for Average-Precision and R-Precision in succession. The average for all query of of Average-Precision and R-Precision is shown in Figure 7. In accordance with result of N1000, the result of N3000 also show that the Concept-based performance is better than Word-based performance according to their precision performances. The improvement of Concept-based precision is 8.3% for average-Precision and 31,5% for R-Precision.

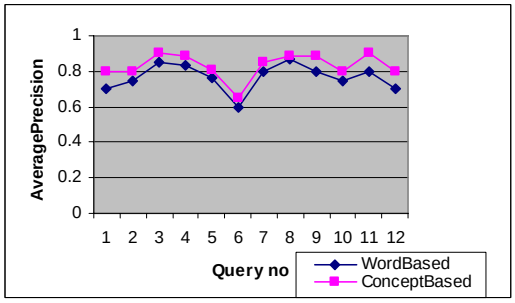


Figure 5. The Average Precision of 12 Query in News3000 test-Collection (Concept-based Beta=0.5,

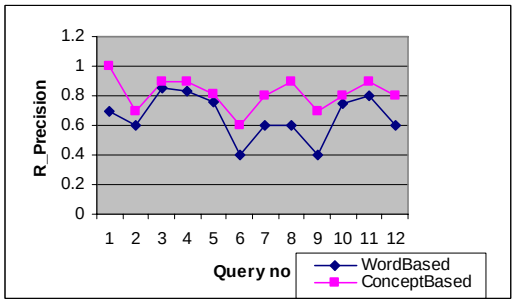


Figure 6. The R-Precision of 12 Query in News3000 test-Collection (Concept-based Beta=0.5,

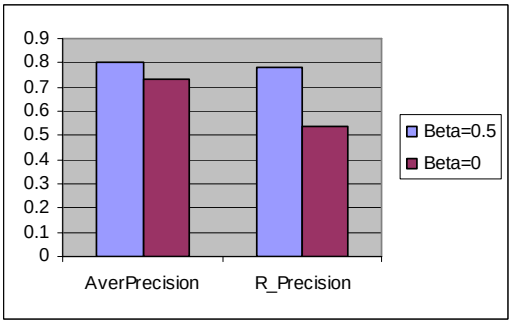


Figure. 7. The Avg- Precision and R-Precision of All Query in N3000 Beta=0.5 (Concept-based) , Beta=0 (Word-based)

*Result for Abstract Collection*

The comparisons between average precision and R-Precision for N3000 test-collection are shown in Figure.8. and Figure.9. for Average-Precision and R-Precision in succession, and Fig 10 for average of average Precision and R-Precision. Although the average precisions for abstract collection are rather lower than news collections, a consistency result that Concept-based is better than Word-based is still recognized.

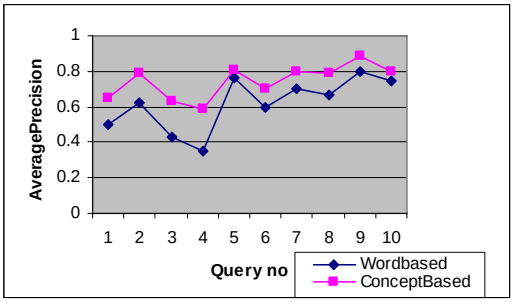


Figure 8. The Avg- Precision for 10 Query in Abstract -collection (CBS=Concept-based Beta=0.5, WBS=Word-based)

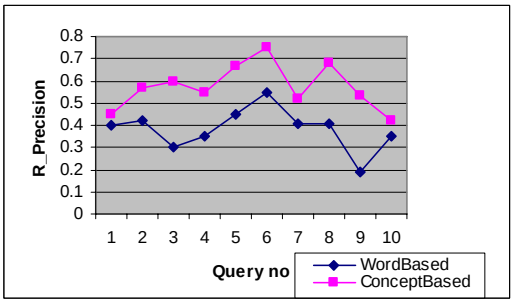


Figure. 9. The R-Precision of 10 Query in Abstract -Collection

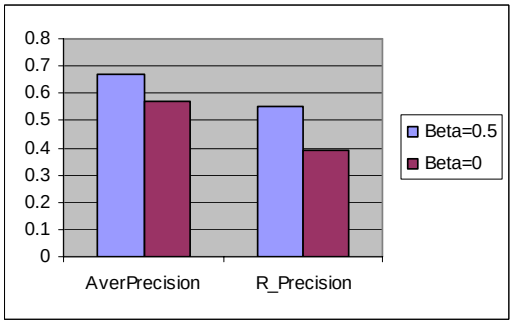


Figure. 10. The Avg- Precision and R-Precision of All Query in N3000 (Beta=0.5 (CBS=Concept-based) , Beta=0 (Word-based)

*The Effect of Beta Setting*

The value of Beta is the weight of Concept to the similarity matching between query and document. If the value is zero the system is purely word-based system, otherwise the system is concept-based according to the value of Beta. In general, the effect of Beta is positive, meaning that the higher the Beta the higher the value of average Precision or R_Precision. For the average Precision the optimal value is Beta=0.9 and for R-Precision the optimal value is Beta=0.5. Although the optimal value of precision is in different Beta value, those Tables show that Beta>0 (Concept-Based) has better precision than Beta=0 (Word-based) for all Beta value (see Figure.11.). Statistical analysis using SPSS with method t-test showed that the difference between Concept-based and Word-based are highly significant for average precision and significant for R-Precision

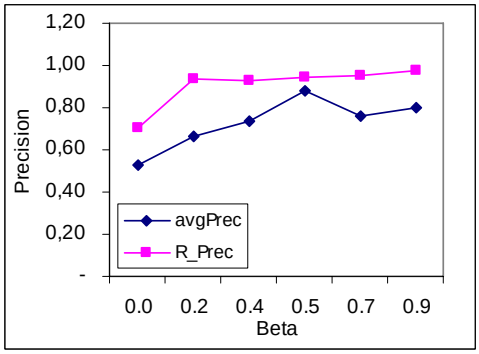


Figure. 11. The Effect of Beta to Average Precision and R- Precision

## CONCLUSION AND FUTURE WORK

Some important conclusions can be drawn from the previous result and discussion. It has been shown that the improvement in precision of Concept-based IR model is statistically significant in average precision and highly significant in R-precision, compare to Word-based IR model. The weighting factor of concept-based is best set to the moderate value around 0.5, that mean same weighting between concept-score and term-score. One of the weakness of this approach is that designer shoud know how many concept are the best for any document collection in the indexing stage.

The suggestion for the future work to improve this method can be mainly focused to the method to predict the number of concept in the document collection efficiently. The prediction of new concept that must be added to the existing concept whenever some new documents is added to the collection is another thing that also important to be consider.

## REFERENCES

Cleary,B. , Exton C., Buckley J. and English M., 2008, *An empirical analysis of information retrieval based concept location techniques in software comprehension*, Empirical Software Engineering Volume 14, Number 1,

93-130, DOI: 10.1007/s10664-008-9095-3.

Egozi, O., 2010,Concept-Based Information Retrieval Using Explicit Semantic Analysis, Master Thesis,Technion Israel Institute of Technology, Heshvan 5770, Haifa.

Egozi,O. , Markovitch, S., and Gabrilovich,E., 2011, *Concept-Based Information Retrieval Using Explicit Semantic Analysis*, Journal ACM Transactions on Information Systems (TOIS), TOIS Homepage archive Volume 29 Issue 2, April 2011

Goyal, P.; Behera, L.; McGinnity, T.M, 2009, *An Information Retrieval Model Based on Automatically Learnt Concept Hierarchies*, IEEE International Conference, ICSC '09, 14-16 Sept. 2009.

Gruber, T., "Toward Principles for Design of Ontologies Used for Knowledge Sharing", International Journal of Human and Computer Studies, 43 (5/6):907-928, 1995.

Guarino, N., "Formal Ontologi and Information System", in N.Guarino (ed), Formal Ontology in Information System, Proc Of the 1st International Conference, Trento, Italy June 1998, IOS Press Amsterdam, pp.3-15., 1998.

Haav, Hele-Mai and Lubi, Tanel-Lauri, 2005, "A Survey of Concept-based Information Retrieval Tools on the Web", Institute of Cybernetics at Tallinn Technical University, Academia Taee 21, 12618 Tallinn.

Jalali V. and Borujerdi, M.R.M. , 2010, *Information retrieval with concept-based pseudo-relevance feedback in MEDLINE*, Knowledge and Information Systems DOI: 10.1007/s10115-010-0327-7

Karypis, G. and Han Eui-Hong, "Concept Indexing A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval and Categorization", Technical Report TR-00-0016, University of Minnesota. www.cs.umn.edu/karypis, 2000.

Rad ,M.P., Hassanpour,H., and Poursaikh, R., 2010, *Concept-Based Information Retrieval with Ontology Approach for Cross-Language Search*, World Applied Science Journal (8): 965-971, 2010, ISSN:1818-4952

Ravindran,D. and S. Gauch,"Exploiting Hierachical Relationships in Conceptual Search", citeseer.ist.psu.edu/ 711765. html, 2004.

Snoek,C.G.M and Worring, M., 2009, *Concept-Based Video Retrieval*, Foundations and Trends in Information Retrieval (2-4)

Steinbach, M., G. Karypis, and V. Kumar , "A Comparison of Document Clustering Techniques", KDD Workshop on Text Mining, 2000.

Van Heijst, G., Shcreiber, A.T., and Wielinga, B.J., "Using Explicit Ontologies in KBS Development", International Journal of Human and Computer Studies, 1997.

Widyantoro,D.H.,2007, *Toward the Development of The Next Generation Search Engine*, Proceeding of The International Conference on Electrical Engineering and Informatics, ICEEI2007, Bandung 17-19 Juni 2007.

Woods,W.A., "Conceptual Indexing : a better way to organize knowledge," Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA, April 1997.

Zeng, J. and Yang, Y., "Information Retrieval Based on Conceptual Network", Internet Research & Development Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China, 2003.

http://www.google.com, 2011