Borrower: IRU

ILL: 18551963 **ILLiad TN: 351248**

Lending String: *IXA,ILU

Patron: Abdelali, Ahmed

Journal Title: ITCC 2005 ; International Conference on Information Technology; Coding and Computing ; proceedings ; 4-6 April, 2005, Las Vegas, Nevada /

Volume: 1 **Issue:**
Month/Year: 2005
Pages: 794-799

03/24/06

Article Title: Rifat Ozcan, Y. Alp Aslandogan; Concept-Based Information Access

Article Author: International Conference on Information Technology; Coding and Computing (6th ; 2005 ; Las Vegas, Ne
Imprint: Los Alamitos, Calif. ; IEEE Computer Soc

Borrowing Notes;

Call #: QA 76.575 I58 2005 V.1
Location: PCL

ARIEL
Charge
Maxcost: rec = $0

Shipping Address:
ILL, NMSU, ZUHL LIBRARY
800-ELP-TAE-TransAmigos Express
1305 Frenger Mall
Box 30003, Dept 3475
Las Cruces, NM 88003

Ariel: 128.123.193.167
Odyssey:128.123.44.152
E-Mail:
Fax: (505) 646-4335

Concept-based Information Access

Rifat Ozcan
Dept. of Computer Science and Engineering
University of Texas at Arlington, U.S.A
ozcan@cse.uta.edu

Y. Alp Aslandogan
Dept. of Computer Science and Engineering
University of Texas at Arlington, U.S.A
alp@cse.uta.edu

Abstract

Concept-based access to information promises important benefits over keyword-based access. One of these benefits is the ability to take advantage of semantic relationships among concepts in finding relevant documents. Another benefit is the elimination of irrelevant documents by identifying conceptual mismatches. Concepts are mental structures. Words and phrases are the linguistic representatives of concepts. Due to the inherent conciseness of natural language, words can represent multiple concepts and different words may represent the same or very similar concepts. Word Sense Disambiguation attempts to resolve this ambiguity using contextual information. The use of an ontology facilitates identification of related concepts and their linguistic representatives given a key concept. Latent semantic analysis, on the other hand, attempts to reveal the hidden conceptual relationships among words and phrases based on linguistic usage patterns. In this work we explore the potential of concept-based information access via these two methods. We examine under what circumstances concept-based access becomes feasible and improves user experience.

1. Introduction

The amount of information that is accessible to an ordinary person today is mind-boggling. While a few centuries ago people were struggling to access information, today many are struggling to eliminate the irrelevant information that reaches them through various channels. The information needs of people are in concept space. Keyword based access to information is sometimes unsatisfactory since it works in word space. Words represent concepts but the mapping from words to concepts is many-to-many. That means one concept may be represented by many different words (synonymy) and one word may represent many different concepts (polysemy). Secondly, since concepts are abstract entities, operational definition and representation of concepts presents a challenge.

In this paper, we present two alternate ways for concept identification: One is based on identifying concepts through Word Sense Disambiguation (WSD) and the second is based on representing concepts through a set of related words using a domain-specific corpus. In the first approach we combined several WSD methods. Then we tested our method on some information retrieval (IR) test collections against traditional word-based indexing. The results show that concept-based IR is more successful on short queries and short documents. Secondly, we identified the concepts in a domain specific corpus using Latent Semantic Analysis (LSA). We tested our approach using the query expansion process, and achieved very encouraging results.

The organization of the paper is as follows: Section 2 describes the challenges of concept-based IR. Section 3 presents our approach to concept-based access. Section 4 discusses the evaluation of our approach. Section 5 encapsulates the related work about concept-based indexing. Section 6 represents concluding remarks.

2. Challenges of Concept-based Access

Figure 1 shows steps in a generic concept-based information access system. In such a system, information content of both information resources and user queries are represented via concepts. Concepts are abstract entities, and various approaches to their representation have been proposed.

0-7695-2315-3/05 $20.00 © 2005 IEEE

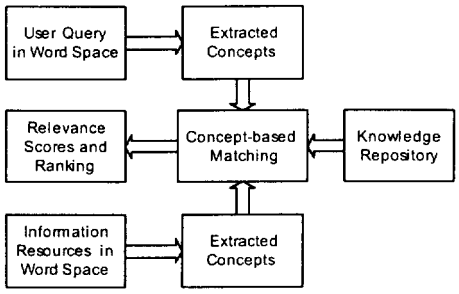

Figure 1: Concept Based IR System

After concept identification and representation, conceptual matching must be done between the extracted concepts with the help of a knowledge repository. The knowledge repository provides information about concepts and their relationships with other concepts in a particular domain. Building such a *practical* knowledge base is another challenge for concept-based access.

3. Approaches to Concept-based Access

We have explored two approaches for concept representation. The first approach represents a concept as a node in an ontology. We use a semantic network representation as in directed labeled graph, which is a simplified conceptual graph [23]. The details of our concept representation can be found in [14]. For the ontology, we are currently using WordNet [10], which can be regarded as a general-domain ontology. A concept is represented as a "synset" that is a set of synonym words that represents the same conceptual entity in the real world, together with its relationships.

Our second approach is the implicit representation of concepts through Latent Semantic Analysis (LSA). Word groups that implicitly represent a concept are identified using a domain-specific corpus. We first present the first approach in detail and then continue with the second approach.

3.1. Concept Identification via Word Sense Disambiguation

We apply Word Sense Disambiguation (WSD) to identify concepts. Our technique is based on evidence combination of supervised and unsupervised WSD methods.

The first method, *Syntalex*, is a supervised WSD system developed by Mohammad and Pedersen [11]. Its feature space consists of bigrams and POS tags of target and surrounding words. The logic behind choosing Syntalex to do evidence combination with our WSD approach is that it uses the local context information and ours focuses on the topical context information.

The second method is based on *WordNet Domains*. WordNet Domains were developed to overcome the widely recognized problem of fine-grained sense discrimination in WordNet [9]. This resource is obtained by tagging each WordNet 1.6 sense with one or more domain labels such as "ARCHITECTURE, SPORT and MEDICINE". The method we use here is a modified version of baseline algorithm defined in [1]. We count the support for each domain in a pre-determined window using the following formula:

$$score_{D(w(i))} = \frac{tfidf}{(i+1) * N_w} \qquad (1)$$

If i^{th} sense of word w is tagged with domain D in WordNet Domains [9], then the score coming from this sense is proportional to the tfidf value of word w and inversely proportional to N_w, number of senses of word w, and the ranking of sense in WordNet. The reason is that words with higher *tfidf* values have much effect on the domains mentioned in a document. On the other hand we penalize the words that has lots of senses and senses that are in low rank according to WordNet ordering because order reflects the frequency of usage of that sense of the word. After this computation, we choose the sense whose domain score is maximum.

The third method is based on the *contextual weights of hypernyms*. This is an unsupervised WSD approach that is similar to the method presented in [2]. It uses hypernym relationships among synsets (concepts) in WordNet. [10]. After finding the named entities in the text, we look up the WordNet dictionary for hypernyms of each possible sense of all words in a specific contextual window. The contextual support for each synset (in the hypernym path of the sense) is calculated using simple frequency calculation. We multiply each contextual weight with decreasing coefficient when we go through the root as given in formula 2. Then the sense that has max score is chosen.

$$Score(s_i) = \sum_{j=0}^{p} c_{ij} * f_{ij} \qquad (2)$$

f_{ij} : contextual support frequency for j^{th} hypernym of sense s_i

c_{ij} : coefficient used for this hypernym

p: The number of hypernyms of sense s_i

3.1.4 Evidence Combination of WSD techniques

We have explored the use of three evidence combination methods: Uncertainty-based, voting, and rank-based.

In the *uncertainty-based* approach, the following uncertainty formula is used to calculate uncertainty of each method that is based on class differentiation quality. [18]

$$H(U) = 1 - \frac{K}{K-1}\sum_{i=1}^{K}(m(i) - \frac{1}{K})^2 \quad (3)$$

K: the number of senses for a word listed in WordNet
m (i): the belief for i^{th} sense of the word. In our case it is the normalized score for each possible sense that is given as output of WSD method.

So we normalize scores for each possible sense. Then uncertainty values are calculated based on the above formula. Then we rely on the WSD technique that has least uncertainty value.

In the case of *voting*, we use the simple voting principle to do the evidence combination in the first phase. That means each WSD method gives its choice for the sense of the word. Then the sense that has maximum vote is chosen as the sense of the word. If there is a tie in the first phase, then we compare the uncertainty values of WSD sources and choose the sense with minimum uncertainty value.

The *sense-ranking* approach is an evidence combination technique considers the sense rankings given by each WSD method. It takes other probable senses into account. The following formula [5] is used to do evidence combination.

$$P(s) = \frac{\sum_k \lambda_k rank_k(s)}{\sum_{s'}\sum_k \lambda_k rank_k(s')} \quad (4)$$

$$rank_k(s) = \left(\left|\{s' \mid P_k(s') > P_k(s)\}\right| + 1\right)^{-1} \quad (5)$$

λ_k is the weight assigned to WSD method k. That shows the reliability of k^{th} WSD method in some way. In our case we used equal weights for each method since they had very similar performances. $rank_k(s)$ is the rank score for sense s given by k^{th} WSD technique. This score is inversely proportional to the number of senses that are strictly more probable than sense s according to k^{th} WSD method output of sense probabilities. Then using the formula 4, we compute sense probabilities for all senses and choose the maximum.

3.4. Concept-based Search with a Domain-specific Corpus and Latent Semantic Analysis Knowledge

Considering the difficulty of WSD techniques and state-of-art results as around 70% precision, we decided to try another way to identify concepts in a document. We assume that only a single sense of a word is relevant to a domain. We used the Infomap-nlp software [8] developed at Stanford University. It takes a corpus as input and produces word pair similarities. A variant of Latent Semantic Analysis is used to reduce the number of dimensions in word vectors [21]. Since we assume that only one sense of a word is relevant for a domain then the output of the program can be considered as related concepts of a concept in the domain. Addition of related concepts to the query helps eliminate the documents that use a different sense of the concept.

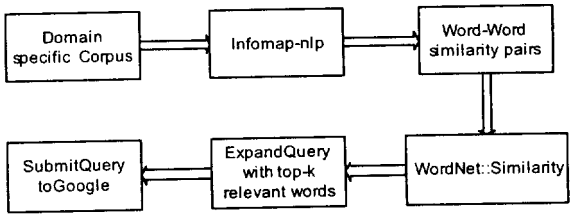

Figure 2: Query Expansion Process

The output of infomap-nlp program [8] contains some words that are not really related to the concept. In [15] the authors present a software called "WordNet::Similarity" that measures the similarity of concepts using several measures. We used this software to compute conceptual similarity of words given by infomap-nlp program. [8] Then we eliminate words that have similarity score that is smaller than some threshold value. The remaining words are used in query expansion process.

4. Evaluation and Results

4.1. WSD Experiments

We did experiments with SENSEVAL-2 English All task data, where documents are sense tagged by human annotators. Choosing the first sense is used as baseline in these experiments. We disambiguated only nouns. Table 1 shows the experiment results.

Table 1. WSD Experiment Results

WSD METHODS	Precision
Baseline (First Sense always)	56.33
CWH	51.64
WSD based on Domain Labels	55.87

Syntalex (Syn)	58.27
Uncertainty (CWH + Syn + B)	56.37
Voting (CWH+Domain+Syn+B)	58.31
Ranking (CWH+Domain+ Syn)	60.00
Ranking (CWH+Domain+ Syn+B)	61.45

CHW refers to WSD based on Contextual Weights of Hypernyms and Domains refers to WSD based on Domain Labels, while B represents the baseline. Considering only nouns, the coverage of our system is 92.1%. Syntalex gave the best result of individual WSD methods. Ranking based evidence combination gave the overall best result by improving the performance of Syntalex by 3.18 %.

4.2. Concept based IR using WSD

We used a dataset that consisted of 2714 image captions and 47 queries [20]. Vector space model is used for indexing with synsets. Figure 3 shows the experiment results. Word based indexing and three variants of concept based indexing interpolated precision values are shown at standard recall points. Synset-based indexing clearly outperformed here. Adding hyponyms and hypernyms of unambiguous words also increased the overall precision values dramatically.

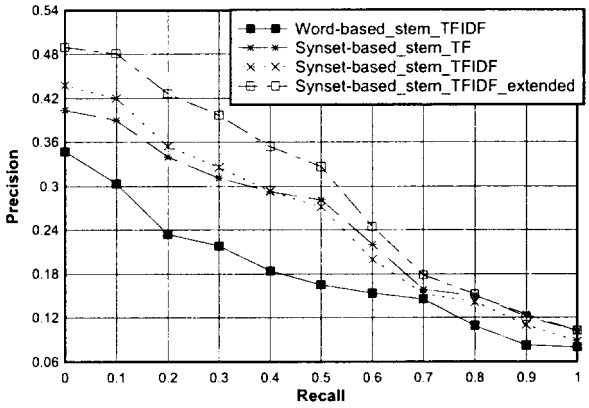

Figure 3. Synset-based Experiment Results

4.3. Concept-based IR based on Latent Semantic Indexing and Query Expansion

We used a part of TEKS (Texas Essential Knowledge and Skills) curriculum as a corpus for infomap-nlp program. As a second corpus we used astronomy category of BankSearch Dataset [18] that contains 1000 html documents. Finally, we constructed another corpus that consisted of 4369 html pages downloaded from Recreational Travel directory of Google [6]. In our experiments, we constructed an expanded query by adding top-5 most related words.

Table 2 shows the precision values for Top-50 documents retrieved for some queries using Google with an original query term and expanded queries. These preliminary results are very encouraging.

Table 2. LSA based Query Expansion Results

Query	Domain	Google	Using LSA
Virus	Biology	6	100
Operation	Mathematics	2	26
Launch	Astronomy	68	98
Star	Astronomy	10	100
Galaxy	Astronomy	45	96
Venus	Astronomy	70	96
Historic home	Travel	88	94
Climbing	Travel	96	100
Reservation	Travel	68	100
Dolphins	Travel	80	92
Average		**53.3**	**90.2**

4.4. Discussion of Results

Firstly, evidence combination techniques improve the performance of individual WSD methods significantly. Secondly, we achieved significant improvement in image captions dataset by concept-based access. In this case word based matching is very poor since queries and documents have less number of words, and a simple matching of words does not suffice. On the other hand, short queries mean less contextual information. This means WSD is more difficult in short queries. The results show that the conceptual information we gained using synsets and especially by adding related concepts through relationships compensate the errors that we have in WSD process due to the lack of contextual information. We also tested our approach on Cranfield test collection but we could not achieve any improvements over traditional approach. The details of these experiments can be found in [14]. Finally, we also tried concept identification through LSA. The results show very significant improvement over original query. However, we need to perform more large-scale experiments on this issue with more queries.

5. Related Work

In [7] detailed survey information on WSD techniques is given. They can be grouped into three: 1) Supervised, 2) Unsupervised and 3) Hybrid methods. Supervised WSD systems [11] generally make use of local contextual information such as local collocation and POS information of surrounding words. Unsupervised methods [12] generally use a lexical resource, such as WordNet and topical contextual information surrounding the target word. They do not require any training data. Hybrid methods [13] use both of these techniques to improve WSD performance.

The effect of WSD on Information Retrieval analyzed by many researches and some contradicting results achieved [17]. Word sense indexing used in [21], but no improvement achieved but even degradation in precision. They concluded that more accurate WSD methods could improve IR. On the other hand, manually disambiguated senses increased IR performance by 29% [5].

LSA is proposed in order to overcome the polysemy and synonym problems of traditional keyword based retrieval [3]. The main goal of this technique is to reveal the underlying semantic structure of the documents by representing them in high dimensional space. LSA uses singular value decomposition in order to reduce the number of dimensions in the term-by-documents matrix. The method is tested in indexing two test corpus [3] and significant improvement achieved.

6. Conclusion

We have explored two methods for concept-based access to information: Word Sense Disambiguation and Ontology-based approach and one using Latent Semantic Analysis. The experimental results shows that we need more accurate WSD systems for long queries-long documents but we showed that even with this accuracy of WSD, conceptual indexing brings significant improvement in short queries-short documents case. The LSA-based approach produced very encouraging results in the query expansion process in travel and education domains. Automatic query expansion mechanisms can be developed to offer this feature when the possible target domains of a user can be estimated with high probability.

Open research problems include exploration of both concept identification methods in semi-automatically built ontologies, and the inclusion of richer relationship types for concept navigation.

7. References

[1] Bernardo M. and Strapparava, C. "Experiments in Word Domain Disambiguation for Parallel Texts", *Proc. of the ACL*, Hong Kong, 2000, pp. 27-33.

[2] Boppana P. and Aslandogan, Y. A., "The 3C Architecture: An XML Topic Maps-Based Framework for Integrating Content, Context and Common Knowledge About Multimedia", *IEEE IRI*, Las Vegas, 2003.

[3] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. "Indexing by latent semantic analysis", *JASIST*, 41(6), 1990, pp.391-407.

[4] Florian, R. Yarowsky, D. "Modeling Consensus: Classifier Combination for Word Sense Disambiguation", *Proc. of EMNLP'02*, Philadelphia, USA, 2002, pp 25-32.

[5] Gonzalo J, Verdejo, F., Chugur I., Cigarran, J. "Indexing with WordNet synsets can improve Text Retrieval", *Proc. of the COLING/ACL '98*, Montreal, 1998

[6] Google Search Engine, http://www.google.com/

[7] Ide, N., Véronis, J. "Word Sense Disambiguation: The State of the Art", *Computational linguistics on Word Sense Disambiguation*, 24:1, 1998, pp. 1-40

[8] Infomap-nlp, http://infomap-nlp.sourceforge.net/

[9] Magnini B., Cavagliá G., "Integrating Subject field Codes into WordNet", *In Proc. of LREC-2000*, Athens, Greece, 2000.

[10] Miller, G.A., Beckwith, R., Felbaum, C., Gross, D., Miller, K. "Introduction to WordNet: An On-line Lexical Database", *Int. Jour. of Lexicography*, 1990, pp.235-244.

[11] Mohammad S., Pedersen T., "Complementarity of Lexical and Simple Syntactic Features: The Syntalex Approach to Senseval-3", *In Proc. of Senseval-3*, Barcelona, Spain, 2003.

[12] Montoyo, A., Palomar, M. "Word Sense Disambiguation with Specification Marks in Unrestricted Texts", *DEXA Workshop*, 2000, pp.103-107.

[13] Montoyo, A., Suarez A. Palomar, M., "Combining supervised-unsupervised methods for Word Sense Disambiguation", *CICLing-2002*. Lecture Notes in Computer Science, México, 2002, pp. 156-164.

[14] Ozcan R. and Aslandogan, Y. A., "Concept-based Information Retrieval Using Ontologies and Latent Semantic Analysis." Technical Report CSE-2004-8, 2004.

[15] Pedersen T, Patwardhan S., Michelizzi, J. "WordNet::Similarity-Measuring the Relatedness of Concepts", in the Proc. of AAAI-04, San Jose, CA, 2004.

[16] Rujie L., Baozong, Y. "A D-S Based Multi-Channel Information Fusion Method Using Classifier's Uncertainty Measurement", Proc. of ICSP2000, 2000.

[17] Sanderson, M. "Retrieving with good sense", *Information Retrieval*, 2000, pp.49—69.

[18] Sinka, M.P., Corne, D.W. "A large benchmark dataset for web document clustering", *Soft Computing Systems: Design, Management and Applications*, 2002, pp. 881-890.

[19] Schutze, H. "Automatic word sense discrimination." *Computational Linguistics*, 1998, pp.97–124.

[20] Smeaton, A.F., Quigley, I. "Experiments on Using Semantic Distances Between Words in Image Caption Retrieval", In Proc. of ACM SIGIR, 1996, pp.174-180.

[21] Voorhees E. "Using WordNet to Disambiguate Word Senses for Text Retrieval", ACM SIGIR, 1993, pp171-180.