A Knowledge-Based Multi-Agent Approach for Initial Query Refinement in Information Retrieval

Tatyana Ivanova[1], Ivan Momtchev[2]

[1] College of Energy and Electronics, Technical University – Sofia, Botevgrad, 2140, Bulgaria

[2] Technical University -Sofia, Blvd. Kl. Ohridski 8, Sofia 1000, Bulgaria

Abstract

Query refinement is one of the main approaches for overcoming natural language lexical ambiguity and improving the quality of search results in Information Retrieval. In this paper we propose a knowledge–rich personalized approach for iterative query reformulation before sending it to search engines and entropy-based approach for refinement quality estimation. The underling hypothesis is that the query reformulation entropy is a valuable characteristic of the refinement quality. We use multi-agent architecture to implement our approach. Experimental results confirm that there is a trend for significant improvement in the quality of search results for large values of entropy.

Keywords: *semantic search, multi-agent system, query refinement, query refinement entropy*

1. Introduction

In general, the quality of returned from search engine results heavily depends on the quality of the sending query. Because of natural language ambiguity (such as polysemy and synonymy) it is very difficult to formulate unambiguous query even for experts in the search domain. That is why techniques to assist the user in perspicuous formulating a search query are needed for improving the performance of the search engines.

Query refinement is a well-known method for improving precision and recall in information retrieval. Query improvement may be made before it is sent to the search engines, or after returning the first results. In the first case the active user involvement and knowledge rich resources are needed. Therefore, this approach is the most effective only for expert users. For non-expert user automatic query expansion, based on statistical analysis of returned results is more successful. In all cases, the quality of the initial query is crucial for obtaining relevant results, since the query reformulation (automatic or iterative) everything is made on the basis of these results. Therefore, knowledge-based methods for query disambiguation are essential for obtaining high-quality-results. Query refinement is a complicated tack, performed in dynamic Web environment and closely depending from unpredictable changing user interests. That is why we believe, that multi-agent architecture is the best one for personalized query refinement.

2. State Of The Art

There are two main research areas, closely related to query disambiguation: Query Refinement (QR, or Query Expansion, QE) and Word Sense Disambiguation (WSD).

2.1. Query Refinement Research

Query Refinement (or expansion) is a process of supplementing a query with additional terms (or general reformulation in some cases), as the initial user query usually is incomplete or inadequate representation of the user's information needs. Query expansion techniques can be classified in two categories (Fig. 1): those based on the retrieved results and those that are based on knowledge. The former group of techniques depends on the search process, uses user relevance feedback in an earlier iteration of search and statistical algorithms (as Local Context Analysis (LCA), Global Analysis [20], and Relevance Feedback techniques) to identify the query expansion terms. Query reformulation is made after initial query sending on the base of the all returned results or using only the first N results automatically (as shown on Fig.2, road (3)) or by little user participation (Relevance feedback, road (4) on the Fig.2). Only the first few of the most frequently used in returned documents terms are used in query refinement process.

Global analysis techniques for query expansion usually select the most frequently used terms in all returned documents. They are based on the association hypothesis,

which states that specific domain terms tend to co-occur in the documents of that domain. Term Clustering, Dimensionality Reduction (LSI, SVD) and Phrasefinder [21] techniques are used for specific term extraction and categorization. While global techniques rely on analysis of a whole collection to discover word relationships, local techniques emphasize analysis of the top-ranked documents retrieved for a query.

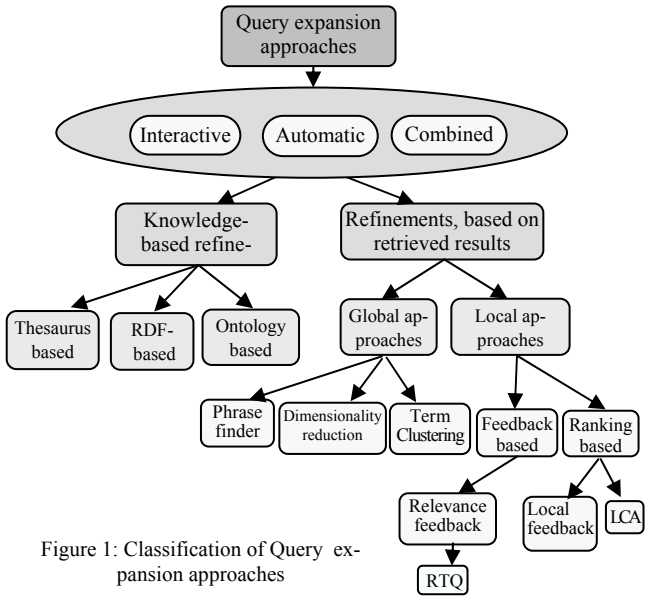


Figure 1: Classification of Query expansion approaches

Local analysis techniques use local selection strategy

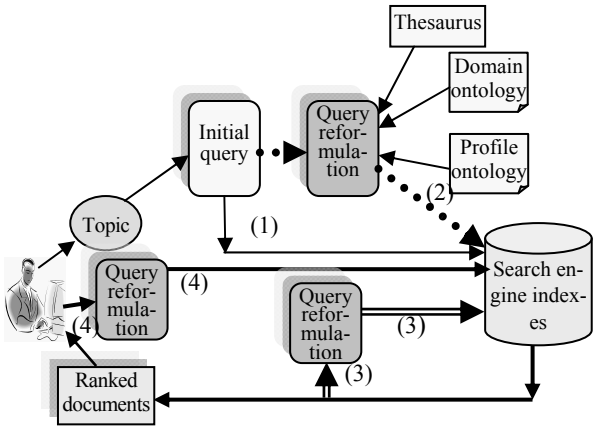


Figure 2: Query refinement paradigms

(usually relation-based) from some group of returned documents. They have shown to be more effective than global techniques when initial query is clear and unambiguous, but they are not robust and can seriously hurt retrieval when few of the retrieved documents are relevant [22]. Experiments on a number of collections, both English and non-English, show that local context analysis offers more effective and consistent retrieval results than global term extraction and categorization techniques [21]. Feedback based approaches use marked from the user as relevant documents for additional term extraction. Relevance feedback [5], [3] is used to refine a query using knowledge of whether documents retrieved by this query are relevant.

The Knowledge-based query refinement approaches [13], [10] are independent of the search process and additional query terms are derived from thesauruses [6], [7] gazetteers, domain ontologies [18], [9], [19], [17] user profile ontologies or other knowledge resources. They use linguistic methods to analyze initial query and provide many possible query expansion terms and rely on the user to select the appropriate ones.

These approaches are usually domain-dependent [18], [6], [9], and are applicable only in domains where the knowledge can be machine-processed. They are applicable in all search stages, but are very useful for initial query refinement, as retrieved document approaches are not applicable in this stage.

2.2. Word Sense Disambiguation

Most Internet searchers tend to use only one or two words in a query [4]. These queries (and often not only they) are unclear and ambiguous. For example the average number of different meanings of common nouns in WordNet is about 7-8. On the other hand, search results are only as good as the posed queries and usually even minimal change in the query leads to significant changes in the results. That is why techniques for word sense disambiguation are of great importance for web search tack.

Word Sense Disambiguation (WSD) is defined as the problem of computationally determining which "sense" of a word is activated by the use of the word in a particular context. The main source of information for the specified word meaning is the context. In general, word sense disambiguation involves the association of a given word in a text with a definition or meaning (sense). The task involves two steps: (1) the determination of all the different senses for the word; and (2) a means to assign each occurrence of a word to the appropriate sense.

There are two main WSD approaches: knowledge-based and corpus-based. Knowledge-based approaches rely primarily on dictionaries, thesauri, lexical knowledge bases, ontologies, without using any textual corpus. They use the definitions from dictionaries, tesauruses, calculate semantic similarity measures using relations from tezauruses or ontologies, or mutual disambiguation algorithms (such as the Lesk method). Corpus-based methods are classification or clustering methods, using tagged or untagged textual corpus and mashine-learning algorithms. These methods are not applicable for initial query disambiguation because of the very little or missing context information in the queries.

3. Our Approach for Interactive Query Refinement

Our experience in Web searching shows that little semantic changes in the query lead to major changes in the returned results. We will call this query property "query volatility". We have made a lot of experiments, sending groups of two semantically very close queries to four search engines: yahoo, hakia, clusty and ask and counting the number of the URLs, returned from the two queries in every group. Some of the results are shown in table 1 for yahoo and hakia, and in table 2 for clusty and ask.

Table 1: differences in the results for similar queries - yahoo and hakia

Search engines / queries	yahoo			hakia		
	All returned results	Identical returned results	Identical returned results (%)	All returned results	Identical returned results	Identical returned results (%)
1) insect 2) insects	1000	88	8,8 %	100	11	11,0 %
1) pelican 2) pelican +bird	1000	31	3,1 %	100	2	2,0 %
1) trigonometry 2) trigonometry + mathematics	1000	233	23,3 %	102	13	12,7 %
1) tiger 2) tiger +wood	998	2	0,2 %	161	0	0,0 %
1) snake 2) snakes	996	135	13,6 %	142	52	36,6 %
1)negative+integer 2) negative+integer +number	988	372	37,7%	100	19	19,0 %
1) dog 2) dog +animal	995	24	2,4 %	100	3	3,0 %
1) abscissa 2) abscissa +axis	994	268	27,0 %	100	9	9,0 %
average			14,5 %			11,7 %

Table 2: differences in the results for similar queries - clusty and ask

Search engines / queries	clusty			ask		
	All returned results	Identical returned results	Identical returned results (%)	All returned results	Identical returned results	Identical returned results (%)
1) insect 2) insects	200	83	41,5%	198	34	17,2%
1) pelican 2) pelican +bird	176	11	6,3%	196	12	6,1%
1) trigonometry 2) trigonometry + mathematics	196	39	19,9%	224	10	4,5%
1) tiger 2) tiger +wood	200	6	3,0%	195	4	2,1%
1) snake 2) snakes	175	84	48,0%	216	30	13,9%
1)negative+integer 2) negtive+integer +number	179	73	40,8%	200	63	31,5%
1) dog 2) dog +animal	198	6	3,0%	198	0	0,0%
1) abscissa 2) abscissa +axis	179	37	20,7%	199	28	14,1%
average			22,9%			11,2%

For example, on the first row in the table 1, number of all results, results, whose URLs are returned from both queries "insect" and "insects" from yahoo, hakia, clusty, ask and their percentage are shown.

The results shows that even when queries are nearly semantically and syntactically the same (as "insect" and "insects" for example) there is significant difference in returned URSs. Because of the query volatility sending as accurate as possible queries is of great importance not only for obtaining good results but for the successful implementation of all query refinement methods, based on the returned results. That is why methods to assist the searcher in query formulation are of great importance for effective search. We propose an approach for interactive query refinement before sending it to search engines.

The conceptual schema of our approach is shown on Fig.3. The approach is knowledge-based and interactive. It heavily relies on the availability of sufficient appropriate sources of knowledge and active participation of the searcher. The process of the initial query refinement is

performed in three stages: (1) Query analysis; (2) Generating query suggestions; (3) Formulating and sending the refined query.

(1). Query analysis. The main aim in this stage is finding the meaning of the query. The system performs syntactic and lexical analysis as searcher type, the query using WordNet. If this analysis fails (WordNet doesn't contain proper nouns, compound domain terms), another knowledge resources as gazetteers, domain ontology or specialized annotated corpus may be used. Many of them may be downloaded from internet (manually or automatically, using "Search web resources module" (Fig. 3).

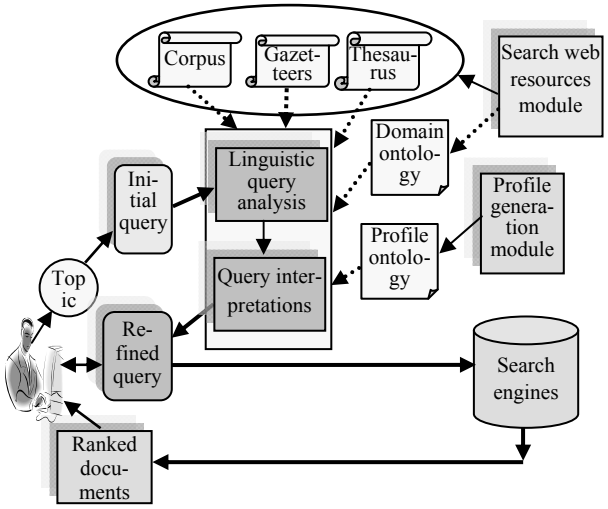


Figure 3. The conceptual schema of the proposed initial query refinement

Search queries usually contain rather keywords than grammatically correct text. This makes linguistic query analysis more difficult and ambiguous than natural language text analysis. For example, when user type only one word, it is often impossible to determine if it is noun or verb, and which among the many possible meanings is in the searcher's account. Using two or more words in the query user rare define clear context and usually the query remains ambiguous. Therefore, even excellent linguistic analysis seldom could find the meaning of the query. That is why a lot of possible query interpretations may be generated as a result of the query analysis.

(2). Generating query suggestions. This stage involves selecting some most appropriate among the possible interpretations and proposing them to the user. The main problem in this stage is making appropriate evaluation of every possible interpretation to restrict the number of proposed interpretations when they are too many. Word sense disambiguation (WSD) techniques are used for restricting possible interpretations. The use of context-based WSD methods is very limited because of missing or incomplete word context in search queries. Reasoning, based on domain ontology, profile ontology (if corresponding system is personalized) or WordNet hierarchy is very useful in this stage.

(3). Formulating and sending the refined query. In this stage the searcher selects an appropriate query reformulation, or formulates the new query if suggestions are inadequate. This is an interactive phase in which the user gives meaning of the suggestions and takes full responsibility for the final query formulation. All the suggestions help him in sending appropriate query, expressing its goals.

The proposed approach helps the user to express more clearly the meaning of the query. The question always arises if adding words, specifying the meaning always came to improving the quality of results. Our experiments have shown that this was not always so. In the next section we propose an approach for estimating the quality of the query refinement, independent of semantics and the subject area of the query.

4. Entropy-Based Approach for Refinement Quality Estimation

Formulating his query, the searcher has in mind a clearly defined semantic. In practice, it appears that the search engines found many semantics, which could lie behind the particular query and returned results are relevant to these semantics. This leads to many irrelevant (for the searcher intense) results, which decreases the precision. It indirectly reduces the actual recall, since some possibly relevant results are shifting from those related to other semantics (and irrelevant in this case). Query refinement leads to narrowing the scope of the search, and thus reducing the total number of hits. Our hypothesis is that there is clear dependence between search engine hits change and quality of semantic query refinement. We assume that the rate of the hits change in the process of query reformulation is the valuable measure for the refinement quality.

Definition: Query refinement entropy (or shortly entropy) is called the quotient of the refined query hits and initial query hits.

Our hypothesis is that the greater is the query refinement entropy, the more significant is the improvement in the quality of the returned results.

Query enrichment should be flexible and able to cope with query ambiguity and react adequately to user information. That is why we propose a multi-agent architecture for implementation of our approach and testing our hypothesis (Fig.4).

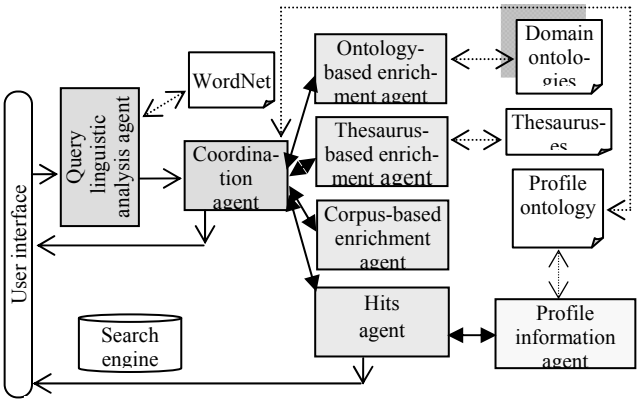


Figure 4 – the multi-agent architecture for query enrichment

The *Query Linguistic Analysis Agent (QLAA)* analyzes the sending by the *User Interface (UI)* query, using WordNet, and sends proposed initial query refinement variants *to Coordination Agent (CA). CA* then proposes various strategies for further query refinement, depending from initial QR proposals, system resources and user profile. It sends request to *Hits Agent (HA)* for delivery hits of existing variants, and asks *Ontology-Based Enrichment Agent* (*OBEA*), *Thesaurus-Based Enrichment Agent (TBEA*) and *Corpus-Based Enrichment Agent (CBEA)* about additional query refinement proposals. They send his proposals back to coordination agent that asks Hits agent for needed from search engine hits, and evaluation of query refinement (using refinement entropy) is returned results to the Coordination agent. *Coordination agent* returns estimated proposals (and its estimations) to the user for choosing the best one. This will help the searcher in clear and unambiguous expression of it information needs. This is only a general ideology of the MAS functionalities. In each real situation the system functioning is determined by the specifics of the query, entered by the user, and can be shown explicitly by using interagent communication diagrams. On the diagram on Fig.5 the query refinement process in case when some of query keywords haven't been found in WordNet is shown; On the diagram on fig.6 the query refinement process by using only WordNet is shown. This may be the whole QR, or the first it phase, when user specify that proposed refinement isn't correct enrichment of his query. In this case, the second QR phase is shown on Fig.7. On Fig. 8 the sequence diagram of QR without using textual corpus is shown, and on Fig 8 – QR when initial information about query domain, represented by ontology is available.
In our subsequent experiments we use mathematical, biological and computer science domain ontologies and don't include thesauruses. Our queries are concept – oriented and implemented in coordination agent strategies are static. This is because of our first aim is to estimate the

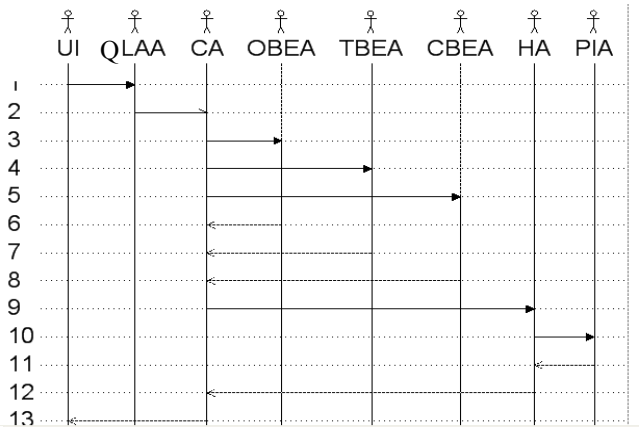


Figure 5 – sequence diagram of QR when some keywords haven't been found in WordNet

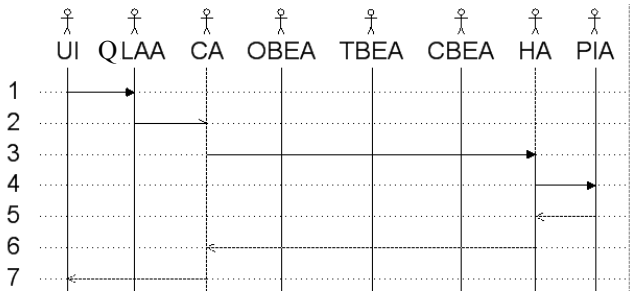


Figure 6 – sequence diagram of QR by using WordNet only

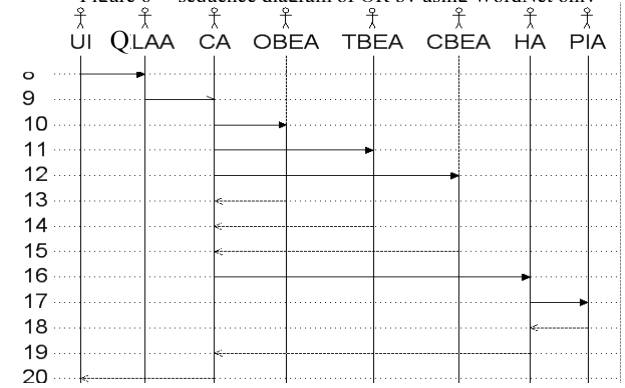


Figure 7 - second QR phase after incorrect enrichment, using WordNet

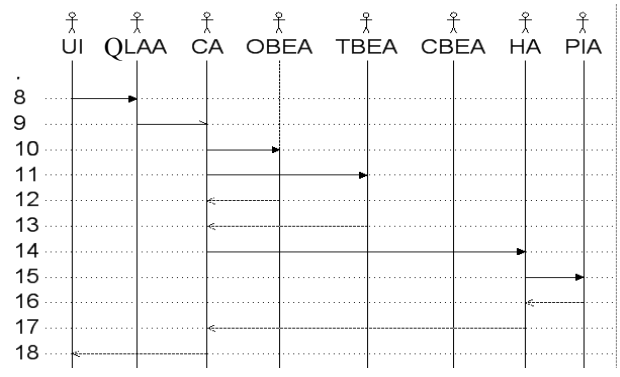


Figure 8 – QR without using textual corpus

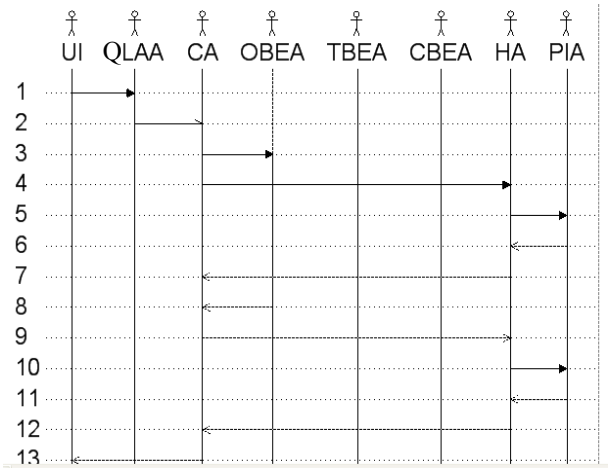


Figure 9 - QR by using WordNet and domain ontology

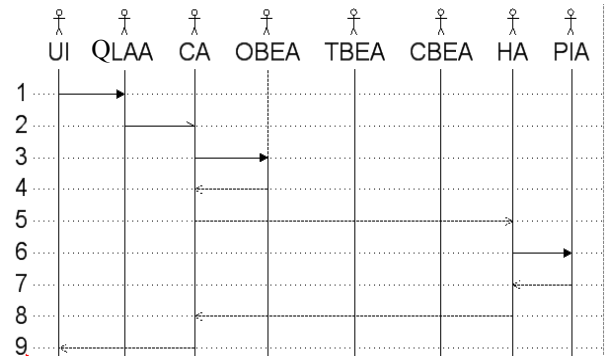


Figure 10 - QR using only domain ontology

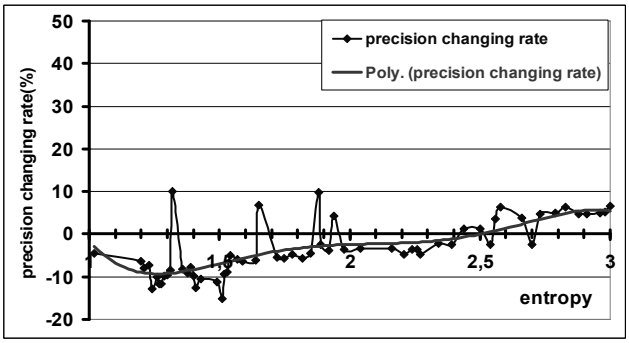
valuability of the proposed entropy-based measure. The full capabilities of the proposed multiagent system are object of future research.

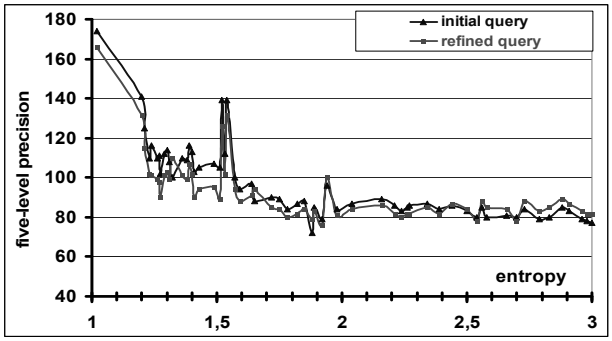
5. Experimental Results

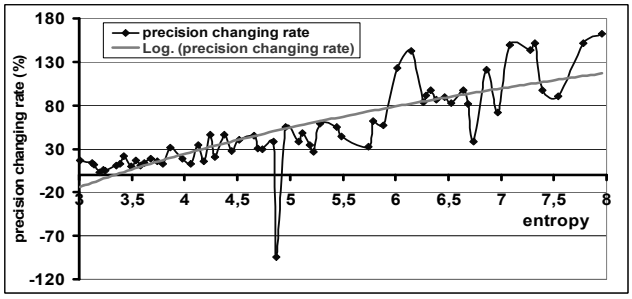
To verify experimentally our hypothesis we have selected 190 search queries, having entropy between 1 and 40400. A key objective in query selection was to provide many different values of entropy, to investigate the relationship between the magnitude of the entropy and returned result's quality. We have manually evaluated relevance of the first 70 results of each sent query and its semantically refined ones by opening and reviewing each of them, received by the original and it semantically refined query. The evaluation is based on adopted common five-level relevance criteria: 4 points for every high-quality result (including high quality relevant to the topic information, represented in a perfect form, including needed hyperlinks to related topics), 3 points for these, lacked some visual representation or links, 2 points for uncompleted information, 1 point for short description of only one of several meanings of the concept, or partially erroneous information, and 0 points for fully irrelevant results. For example, when we send query "tiger", searching information about an animal, we assign:

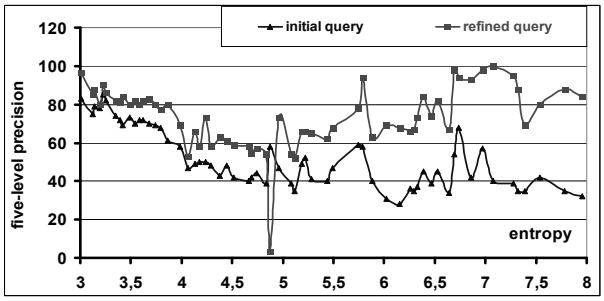
- 4 points to wikipedia's comprehensive hyper-linked and illustrative material http://en.wikipedia.org/wiki/Tiger;
- 3 points to http://www.knowledgerush.com/kr/encyclopedia/Tiger/;
- 2 points to http://www.bbc.co.uk/nature/ species/Tiger;
- 1 point to http://wordnetweb.princeton.edu/perl/webwn?s=tiger;
- 0 points to: http://www.apple.com/macosx/; http://www.bbc.co.uk/nature/species/Tiger; http://sports.yahoo.com/golf/pga/players/147, http://www.jobtiger.bg/, as fully irrelevant to the animal's topic.

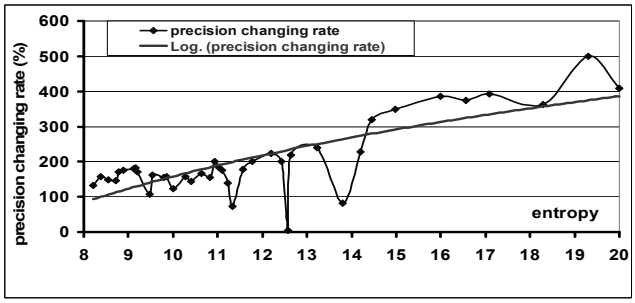
We group the results according to the entropy value in five groups (for achieving clear visual representation): first group with entropy between 1 and 3 (first two diagrams on Fig. 11 and Fig. 12), second group having entropy, greater than 3 and non greater than 8 (second two diagrams on Fig. 11 and Fig. 12), third group having entropy, greater than 8 and non greater than 20 (third two diagrams on Fig. 11), forth group, having entropy, greater than 20 and non greater than 100 (forth two diagrams on Fig. 11 and Fig. 12), and the last group, having entropy, greater than 100 (the last two diagrams on Fig. 11 and Fig. 12).We call the sum of the points of the first 70 received results of the initial and refined query five-level-precision and show it values on the left group of charts. On the right group of charts we show the dependency between entropy and five-level-precision rates of change, as well as polynomial or logarithmic approximation of this dependency. We think five-level-precision is a good results evaluation measure, because usually relevant results are of varying quality and no one of frequently used result's relevance measures (precision, recall, f-measure and so on) takes note of this fact

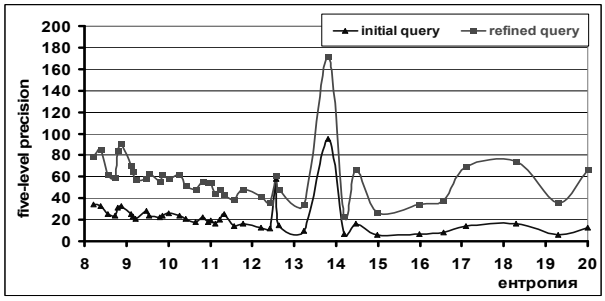


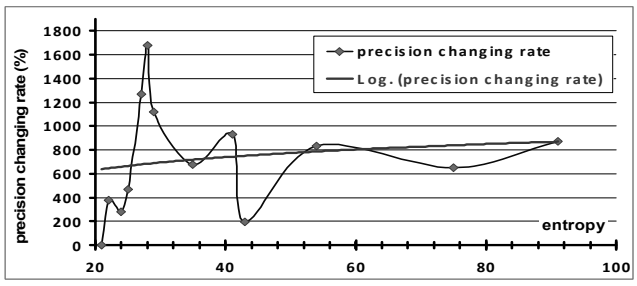


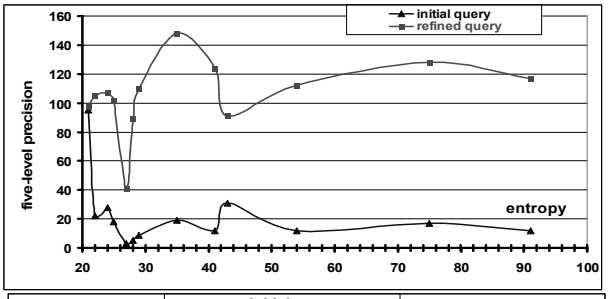


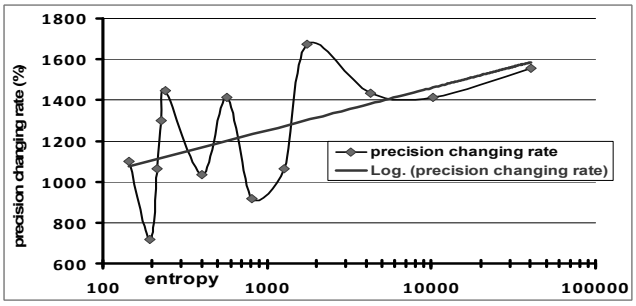












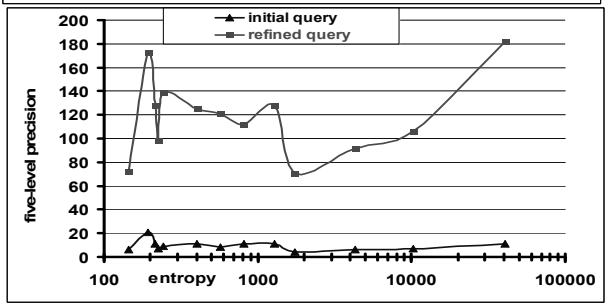


Figure 11. Graphical representation of experimentally evaluated dependency between precision and entropy

Figure 12. Graphical representation of experimentally evaluated dependency between precision and entropy

6. Conclusion

In this paper we propose a knowledge–rich approach for interactive query disambiguation before sending it to search engines and entropy-based approach for refinement quality estimation. Our studies are aimed at exploratory search of textual web resources. The literature review has shown that a large variety of knowledge-based sources and methods may be used to help the user in the web query sense disambiguation. To achieve the best results combined use of many of them is needed. We propose a multiagent architecture for flexible dynamic combination of various query analyzing and disambiguation methods and algorithms. Having in mind the prevailing web search engine's syntactic keyword-based searching and indexing approaches, our main goal was to find the answer of the question whether (always) the semantic query refinement leads to better search results. For our experiments we implement above approaches, using Jade-based Multi-agent system. The experimental results have shown that semantic query refinement not always leads to results improvement. When the entropy is not larger than 3, returned from the refined query relevant results in the group of first 70 usually are more, but some of the very good ones (returned by the initial query) are no longer among the first 70 and as a result the overall quality of the results in many cases is not better. With the increase of entropy the expectations for better results are greater, but even for large values of entropy there are isolated cases of worsening quality of results. There is a trend for significant improvement in the quality of search results for large values of entropy.

Despite the clear tendency to increase the quality of results as a result of the increasing entropy, there is a large deviation from this trend in both directions. This means that in case when entropies are slightly different each-other, in some cases query disambiguation leads to much better results than other ones. Our future research will be directed to search other dependencies (apart from the values of entropy) between the query refinement and the quality of returned results and experiment various strategies for query disambiguation, using various knowledge-based resources.

As experiments show, query entropy is a valuable measure for query-refinement quality and it should be used for calculating of expected refinement effect of various manually or automatically added disambiguation words. Complexity and indeterminism of query disambiguation methods, as well as unpredictability and dynamism of web environment and user needs make appropriate use of multiagent architecture for implementing query-refinement tack. The query entropy is one of the valuable criteria for choosing of appropriate query refinement interactively from the user or automatically from Query enrichment evaluation agent. Practical realization of query refinement heavily depends from coordination agent's strategies. These strategies are important field of future research.

References

[1] E. Agirre, et al., "Semeval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems", 2007, http://acl.ldc.upenn.edu/W/W07/W07-2002.pdf

[2] E. Agirre, P. Edmonds, "Word Sense Disambiguation: Algorithms and Applications", 2007

[3] N.J. Belkin, et al., "Relevance Feedback *versus* Local Context Analysis as Term Suggestion Devices", Rutgers' TREC−8 Interactive Track, 2000

[4] M. *K.* Bergman, "Guide to Effective Searching of the Internet", tutorial, *2005*

[5] B. Billerbeck, and J. Zobel, "Techniques for Efficient Query Expansion", 2004, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.8466&rep=rep1&type=pdf

[6] D. Buscaldi, et al., "A WordNet-based Query Expansion method for Geographical Information Retrieval", 2005, http://clef-campaign.org/2005/working_notes/workingnotes2005/buscaldi05.pdf.

[7] P.A. Chirita, et al., " Personalized Query Expansion for the Web", in proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007

[8] H. Dang, "The Role of Semantic Roles in Disambiguating Verb Senses", 2005, http://acl.ldc.upenn.edu/P/p05/P05-1006.pdf

[9] F. Gaihua, et al., "Ontology-based Spatial Query Expansion in Information Retrieval", 2006, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.940&rep=rep1&type=pdf

[10] F.A. Grootjen, "Conceptual query expansion", Data & Knowledge Engineering, 2006, http://www.cs.mun.ca/~hoeber/download/2005_awic.pdf

[11] B. Katz, et al., "Word Sense Disambiguation for Information Retrieval", 1999, http://aaaipress.org/Papers/AAAI/1999/AAAI99-192.pdf

[12] S. Liu, "Word Sense Disambiguation in Queries", 2005, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.973&rep=rep1&type=pdf

[13] A.P. Natsev, et al., "Semantic Concept-Based Query Expansion and Re-ranking for Multimedia Retrieval. A Comparative Review and New Approaches", 2007, www.aquaphoenix.com/publications/mm2007_semantics.pdf

[14] Pitler, E., et al., (2009), "Using Word-Sense Disambiguation Methods to Classify Web Queries by Intent", http://140.116.245.248/ACL-IJCNLP-2009/EMNLP/pdf/EMNLP148.pdf

[15] W. Ryen, G. Marchionini, "Examining the Effectiveness of Real-Time Query Expansion", 2009, http://cchen1.csie.ntust.edu.tw/students/2009/Examining%20the%20Effectiveness%20of%20Real-Time%20Query%20Expansion.pdf

[16] A. Sanfilippo, et al., "Word Domain Disambiguation via Word Sense Disambiguation", 2006, http://acl.ldc.upenn.edu/N/n06/N06-2036.pdf

[17] G. Solskinnsbak, "Ontology-Driven Query Reformulation in Semantic Search", 2007, http://daim.idi.ntnu.no/masteroppgaver/IME/IDI/2007/3523/masteroppgave.pdf

[18] N. Stokes, et al., "Exploring criteria for successful query expansion in the genomic domain", 2009, http://csserver.ucd.ie/~nstokes/publications/nstokes_IR_SI_TREC09.pdf

[19] S. Tomassen, "Research on Ontology-Driven Information Retrieval", 2006, http://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.106.4952&rep=rep1&type=pdf
[20] J. Xu, and W. B. Croft., "Query expansion using local and global document analysis", In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, 1996, http://www.iro.umontreal.ca/~nie/IFT6255/xu-croft.pdf
[21] J. Xu, and W.B. Croft, "Improving the Effectiveness of Information Retrieval with Local Context Analysis", ACM Transactions on Information Systems, 2000 http://www.ir.iit.edu/~dagr/IRCourse/Fall2000/Presentation/ OriginalPapers/TOISLocalContextAnalysis.pdf
[22] X. Xu, "A Comparison of Local Analysis, Global Analysis and Ontology-based Query Expansion Strategies for Biomedical Literature Search", 2006, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.6402&rep=rep1&type=pdf

D-r Tatyana Ivanova, 2009, is an assistant-professor in College of Energy and Electronics (CEE), Technical University of Sofia, from 2000.

Assoc prof. d-r Ivan Momtchev is a dean of French Language Faculty of Electrical Engineering, Technical University of Sofia, from 2005.