

# Learning Contextual Behavior of Text Data

*Coskun Bayrak*

*Department of Computer Science  
University of Arkansas at Little Rock,  
Little Rock, Arkansas 72204 USA  
[cbayrak@ualr.edu](mailto:cbayrak@ualr.edu)*

*Hemant Joshi*

*Department of Applied Science  
University of Arkansas at Little Rock,  
Little Rock, Arkansas 72204 USA  
[hmjoshi@ualr.edu](mailto:hmjoshi@ualr.edu)*

## Abstract

*Understanding contextual behavior is very important in order to develop a context-aware retrieval system. This paper discusses the philosophy behind the development of the “Evolutionary Behavior Of Textual Semantics” (EBOTS) system. The representative mathematical model has complemented the theoretical foundation. Intuitive contextual behavior is studied as a part of proposed research work. Context retrieval based on semantic knowledge allows generic queries to be defined, instead of exact word-based queries. The results of the context retrieval for a classic3 and Time dataset using the EBOTS system have been discussed in this paper. The paper makes a contribution to the semantic knowledge representation and retrieval algorithms.*

## 1. Introduction

The World Wide Web contains 170 Terabytes of information [1] and is growing at a fast rate. Most of the new information produced is in the digital media, ensuring lower cost and a better outreach. Text is the most popular medium to express descriptive stories, novels, scientific publications, reports, news and emails, etc. Text analysis is a promising field with newly available tools in Natural Language Processing (NLP), Machine Learning, and Information Retrieval (IR). Different languages are used to express human thoughts. Different languages sustain different literary styles that are not exactly a scientific phenomenon. No matter what language is used, language is a lossy compression of human thoughts [2].

Characters contribute to form meaningful words and words have basic inalienable meanings [3]. A homogeneous set (from the same language) of the words in a particular order following grammatical rules gives rise to a sentence. Even though words have different meanings, they express one intended meaning (by writer or orator) within the given sentence boundary [4]. Thus, different words, when they occur with one another in a certain order; give rise to a context. Sentences convey contextual meaning. The study in this paper is intended to analyze the contextual behavior of text media.

So, how is the context different from a concept? Concepts are more generic and abstract in their ability to convey a particular message. Context boundaries are small and refer to an expression of the thought. On the other

hand, the concept boundaries cannot always be defined, due to their abstract nature. A given concept can be expressed with a few words, a paragraph or an entire book.

In order to study the contextual behavior in languages like English, focus has to be on the sentences and their boundaries. Better knowledge management can be achieved by preserving the original text structure in which literature has been presented. The hierarchical representation of language allows the EBOTS architecture to maintain a meaningful correlation between the words. It also allows the EBOTS system to extend the available knowledge representation into various useful forms. The knowledge representation forms may include ontologies for qualitative and semantic representations as well as the term-document structure followed by statistical IR methods.

Semantic context representation is significant in the context-based retrieval systems. Contextual Retrieval is a semantic subset of traditional IR. Typical IR system primarily focuses on the accuracy of the results by increasing precision and recall [5]. In the case of Context Retrieval systems [6], the focus is not only on the precision and recall but also on abstract query context retrieval. The query context can be generic or specific to a particular field for which retrieval can be performed [7].

The methodology behind the EBOTS system is explained in detail in the following sections. In section 2, the basic formalism behind the context aware framework of the EBOTS system is presented. Section 3 will discuss a formal model for context modeling. Section 4 primarily discusses experiments performed and results obtained. Finally, Section 5 will consider the conclusion and the future enhancements that can be contributed.

## 2. Background

The EBOTS system will identify the context of sentences in the given text by using the meanings of different words in that context to represent acquired knowledge. Once the words are identified in a particular context, the system forms the hierarchy of words in a tree-like structure. Each tree represents the direct context of a single sentence. Each tree can be formed for a sentence and then extended as these sentences belong to document(s). These trees can even form paragraphs and documents when extended.

In Information Retrieval techniques, Term Frequency (TF) and Inverse Document Frequency (IDF) are commonly used as local and global weights respectively. Normalization along with local and global weights determines total weight of each term in the dataset. Filtering techniques are also widely used to remove less frequent terms or too common terms. The weighing techniques are important aspect for determining which words are important and which words are less significant.

The theoretical foundation of the EBOTS system is composed of reference domains, and correlation types (strong, weak and/or no correlation) between the reference domains. A brief definition of each of these follows.

### 2.1. Reference Domain

A reference domain can be defined as the group of words of a single sentence that represents similar contextual properties. A Reference domain represents the context of text data. Reference domains follow sentence boundaries. The words form the elements of reference domains. When two reference domains have common element(s), they are more likely to share the same context. Hence the correlations can be defined among the related reference domains. These correlations can be of three different types.

Reference domains consist not only of synonyms [8] but they also include contextual information for every word in the sentence. Consider a simple sentence like "Mary had a little lamb". Only three words, namely *Mary*, *little* and *lamb* are considered after filtering common words and stop words to find word roots [9]. The resulting structure is shown in Figure 1. The EBOTS system uses a Machine Readable Dictionary [8] to obtain synonyms and glossary meanings of each word.

**2.1.1. Strong Correlation.** A single document consists of a number of reference domains. If an element 'a' of one reference domain A is common with the element of reference domain B, and if A, B belong to the same document, *Strong* correlation is said to exist among the two reference domains. This is logical because if the two sentences appear in the same document, and if both of the sentences share the common (partial) context, then both of the sentences are likely to share same meaning as well.

**2.1.2. Weak Correlation.** Consider the case in which two documents consist of two reference domains A and B respectively. If these two reference domains share one or more elements, there is a Weak correlation between the two reference domains. When two documents share common elements among their reference domains, they are likely to have partial common context.

Sharing of the common context is necessary but not the sufficient condition to determine the correlation between the two documents. If two documents have very similar content, they are likely to be weakly correlated. Especially the duplicate documents will have weak correlations

readily established. There is a level of fuzziness involved whether having common elements in two documents is a sufficient condition for the two documents to have same meaning. This fuzziness is well represented with a weak correlation.

| Mary had a little lamb |                                      |   |
|------------------------|--------------------------------------|---|
| mary                   | Virgin Mary<br>Madonna<br>The Virgin | the mother of Jesus; Christians refer to her as the Virgin Mary; she is especially honored by Roman Catholics |
| little                 | fiddling                             | small and of little importance  |
|                        | brief                                | of short duration or distance   |
|                        | lilliputian                          | Small in size   |
|                        | younger                              | Younger by age  |
|                        | miniscule                            | Small in size and or shape  |
|                        | little                               | small in a way that arouses feelings  |
| lamb                   | lamb                                 | Younger sheep   |
|                        | Charles Lamb, Elia                   | English Essayist  |
|                        | lamb                                 | a person easily deceived or cheated   |
|                        | dear                                 | a sweet innocent mild-mannered person   |
|                        | lamb                                 | the flesh of a young domestic sheep eaten as food   |
|                        | lamb                                 | give birth to a lamb  |

**Figure 1** Detailed Reference Domain structure for a simple sentence "Mary had a little lamb"

**2.1.3. No Correlation.** If there is no common element between the two reference domains, nothing can be predicted about whether they share the same context or not. Irrespective of whether the two reference domains exist in the same or different documents, no commitment can be made about their correlation and *No* correlation is said to exist in this case.

## 3. Formal Model of Correlation

Theoretically, correlations are helpful in determining direct or indirect association with a particular concept, but it is important to measure its quality mathematically. Based on the correlation theory, a formal model is presented.

Consider  $T$  to be the total number of correlations that are referenced to a particular reference domain,  $A$ . Out of the total  $T$  correlations  $T_w$  is the number of the weak correlations of other reference domains with  $A$  while  $T_s$  is the number of strong correlations that are established by other reference domains with  $A$ .

$$T = T_s + T_w \quad (1)$$

### 3.1 Correlation Delta

The correlation delta of any given reference domain,  $A$  is a formal representation of its significance to the other reference domains in the system as well as its relevance to the query context. The correlation delta is computed for every reference domain in the EBOTS system with respect to the given query context. It is also a useful representation to find out how many reference domains share a common context. Correlation delta  $\Delta_i$ , for reference domain  $i$ , is defined as:

$$\Delta_i = \left\lceil \frac{T}{I+T} \right\rceil * \left( \frac{I+T_s}{I+T} \right) \quad (2)$$

The equation (2) achieves normalized correlation delta values in the range between 0 and 1 (inclusive),

i.e.  $0 \leq \Delta_i \leq 1$

The formula in equation (2) is significant in determining the strength or weakness of the reference domain with respect to the query context. The formula in equation (2) can be considered as split into two parts. The first part uses a ceiling function (indicated by  $\lceil \text{and} \rceil$ ) that will always round up the division of total links, T and its divisor to the next higher integer value. The reason number of total link is divided by (T+1) is because if there are no links available for the given reference domain, division by zero situation is avoided. But intuitively speaking, the number of total links available is calculated from outside references to that particular reference domain. Additional correlation i.e. T+1 is established by the single self referencing link of the reference domain to itself. The second part of the formula ensures that only if the number of strong correlations is equal to the number of total correlations ( $T=T_s$ ), then the value of second part will be 1, else it will be some fraction value (between 0 and 1) indicating the ratio and fuzziness of relevance.

The correlation delta value of zero refers to an isolated reference domain ( $T=0$ ) that does not share a context with any other reference domain of the system. Correlation delta value of 1 indicates presence of strong correlations for that particular reference domain. High correlation delta value 1 is possible only when all correlations with the reference domain under consideration are strong ( $T=T_s$  and  $T_w=0$ ).

As the number of weak correlations increases, the strength of its correlation to the query context decreases. In this case, the value of the correlation delta lies between 0 and 1 (excluding 0 and 1) i.e.  $0 < \Delta_i < 1$ . It should also be noted that the correlation delta is directly proportional to the number of strong correlations and inversely proportional to the number of weak correlations. The weak correlation delta values allow fuzziness of correlation strength among reference domains for a particular query context.

### 3.2 Example

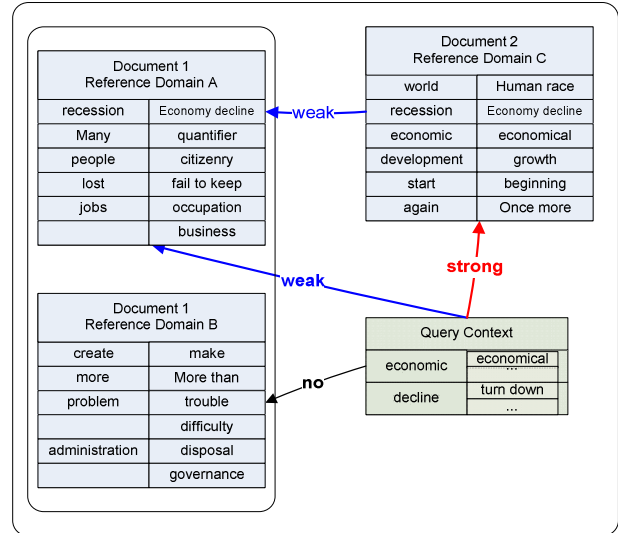
Consider the example shown in Table 1, which consists of 4 documents. Example of 4 different documents is shown.

**Table 1** Four sample documents with different subjects.

| Document 1   |
|--|
| Due to <u>recession</u> <u>many</u> <u>people</u> have <u>lost</u> their <u>jobs</u> . |
| This <u>creates</u> <u>more</u> <u>problems</u> for the <u>administration</u> .        |

| Document 2   |
|--|
| In 1931 <u>world</u> <u>recession</u> was over and <u>economic</u> <u>development</u> started <u>again</u> . |
| Document 3   |
| The <u>US</u> <u>economy</u> has been <u>hit</u> <u>severely</u> in <u>recent</u> <u>years</u> .             |
| The <u>budget</u> <u>deficit</u> is the <u>highest</u> in the <u>last</u> <u>50</u> <u>years</u> .           |
| Document 4   |
| <u>Central</u> <u>park</u> is a <u>beautiful</u> <u>place</u> in <u>New</u> <u>York</u> .                    |

The analysis is started by forming the reference domains. For simplicity, initially only documents 1 and 2 are analyzed. Notice that common stop-words were filtered in the formation of reference domains. Also notice that the reference domains A and B belong to the same document. Query context *economic decline* will be used for determining correlations between the reference domains. This can be represented in terms of reference domains as shown in Figure 2.



**Figure 2** Reference domains of the EBOTS system and representation correlations among them for the context *economic decline*

The correlation between the query context ‘*economic decline*’ and the reference domain C (of document 2) is considered as *strong*, because the reference domain C contains the partial contextual reference to the query. The correlation between the reference domain A and the query context should be considered weak. This is justified because semantically *economic decline* is similar to recession and reference domains A and C share a common element i.e. *recession*. The strength of the argument here is that both the reference domains A and C share a common context and the two reference domains are likely to be similar. As the two reference domains belong to two different documents, the correlation type between them is *weak*. Finally the reference domain B in document 1 does not share a common context and has *no* correlation to the query context.

In the example shown in Figure 2, document 1 is about the recession and problems due to it. Document 1 does have a partially common context with document 2 but this does not guarantee that they are exactly similar. Document 2 is about historical perspective of the world recession during 1931. Document 3, which consists of 2 sentences details about the US economy and the budget deficit.

If another document is introduced with a sentence like “*Central park is a beautiful place in New York*” (Document 4), then *No* correlation can exist between this new document and any of the existing documents or the query context.

### 3.3 Document Correlation

As a hierarchical representation, a document repository consists of many documents. Each document consists of one or more reference domains. Each reference domain consists of a set of words that help to describe the context. The TF and IDF values are calculated for each unique word in the dataset. The average TF and average IDF can be calculated for each sentence (or reference domain) using TF and IDF values of each word in the sentence. Also normalization can be applied to average TF \* IDF values in order to obtain normalized average TF \* IDF values of a single reference domain. The average TF \* IDF values for each reference domain are necessary but not sufficient in determining relevance to the query reference domain(s). Thus, normalized average TF \* IDF values are combined with correlation delta as the weight bias factor towards relevance. These weights can be used in two different ways to retrieve relevant documents. Both the approaches are discussed.

#### 3.3.1 Vector Space Model (VSM) approach

In the VSM approach, the reference domain vs. document matrix is formed. The correlation delta can be defined for the reference domains using equation (2). In the VSM approach, the correlation delta of each sentence acts as weight along with normalized average TF and average IDF weights of a sentence to form the total weight. In essence, this offers sentence by document representation instead of traditional term by document representations. The normalization factor  $N_k$  can be used for the reference domain  $k$  with correlation delta  $\Delta_k$ . Normalization process ensures that the sentences that are longer in length (in terms of number of words), do not influence the total weight of the sentence. The contextual properties are preserved within the correlation delta calculated for strength of a particular sentence to the query concept. So for sentence  $k$ , the total weight is

$$Total_{weight(k)} = Avg.TF_k * Avg.IDF_k * N_k * \Delta_k \quad (3)$$

Here,  $Avg.TF_k$  indicates normalized average TF value for the reference domain  $k$  and  $Avg.IDF_k$  is the normalized

average IDF value of the reference domain  $k$ . Using equation (3) reference domains versus document matrix, queries can be semantically mapped to the vector space to find the similarity.

#### 3.3.2 Document Delta Approach

Alternate method is to start with correlation delta for each reference domain and in bottom-up manner go on calculating average summation of correlation delta for each document. The document delta  $\nabla$  represents an average of individual correlation delta values of all reference domains in that document. Documents with high document delta values represent a strong coherent association to the context. Documents are organized by descending order of document delta values to determine their relevance to the query. For a particular document  $d$  with  $N$  reference domains, document delta  $\nabla_d$  is defined as:

$$\nabla_d = \frac{\sum_{i=1}^N \Delta_i}{N} \quad (4)$$

Using the equation (3), the overall weight of each reference domain can be calculated for any given query context. With the weight factor available, two different approaches can be deployed. In the first approach, reference domain vs. document matrix is formed and query context is mapped to the vector space to obtain similarity between query and available documents. In the second approach, only equation (4) is used to calculate relevance of each document with the query context. Documents with higher values are more similar than others.

## 4. Experiments and Results

Using the aforementioned approach, several experiments were conducted with the classic3 [10] dataset and Time magazine news articles dataset. The context retrieval results obtained from the EBOTS system were analyzed and compared with those resulting from the use of other common methods.

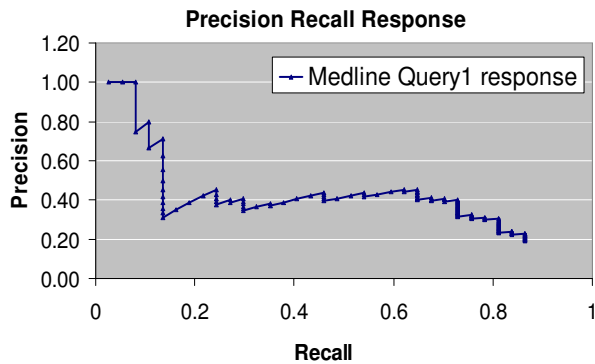
Normally, TF and IDF are determined for every unique term in the corpus [12]. The Term vs. Document matrix is formed with terms as rows and documents representing columns. The values represent the product of TF and IDF values of a particular term. Query term(s) are also mapped to a virtual query document vector. The dot product of the column vector with the query vector produces the cosine of angle between the two vectors [11]. Various weighing schemes for local, global weighing of the terms have been suggested [12]. Typical Vector Space Model employs term weighing techniques in order to determine the similarity between two documents.

As mentioned earlier, two approaches were used to determine sentence based query relevance. Using document delta approach is more simplistic and intuitive but is susceptible to noise present in the data. On the other hand, forming reference domain by document matrix may be computationally intensive but provides gradual relevance judgment for each document in terms of angle of similarity. In the following section the results of the experiments are presented.

#### 4.1. Classic3 Dataset

The Classic3 data corpus [10] consists of 3 datasets namely MEDLINE, CISI and CRAN. The MEDLINE dataset consists of 1033 abstracts from medical journals, CISI consists of 1460 abstracts from information retrieval field, and CRANFIELD consists of 1400 abstracts from aeronautical systems area. Figure 3 shows the Precision-Recall response of the EBOTS system for the MEDLINE collection query 1 using VSM. The vector space model was formed using reference domain vs. document matrix of values obtained by equation (3). Mean average precision for MEDLINE query 1 was 0.998. Several experiments were conducted with different queries and different dataset such as CISI and CRANFIELD.

The EBOTS system achieves comparable results with a reduced matrix size. In contrast to traditional approaches, which use unique words to form rows, the EBOTS system employs techniques to represent entire sentence in each row. With the (Latent Semantic Indexing) LSI technique, a reduced rank semantic representation can be achieved but the abstract queries can achieve better results with the simplified EBOTS matrix representation.

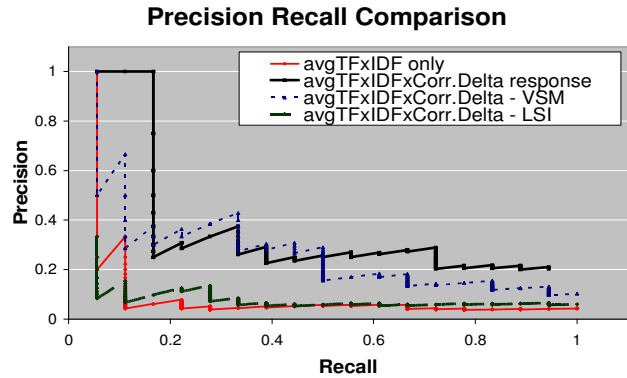


**Figure 3** Precision-Recall response of the EBOTS system for the query 1 of MEDLINE dataset

Correlation delta threshold of 0.75 was used for MEDLINE dataset i.e. Sentences with correlation delta more than 0.75 were selected as candidate. Once the candidate documents are selected based on weighing criteria, they are ranked in descending order of their similarity.

In Figure 4, precision-recall response is shown for first query of Time world news collection from 1963 Time magazine. This dataset consists of 424 articles. The

experiments showed 100% precision at 18% recall for query 46 provided with the dataset using document delta approach. For time magazine dataset, correlation delta threshold of 0.6 was used.



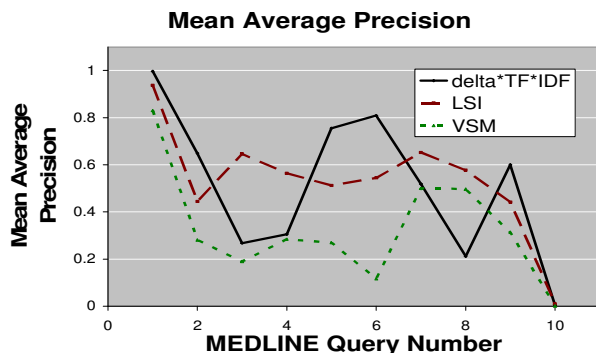
**Figure 4** Precision-Recall response of the EBOTS system for the query 46 of Time dataset

The candidate documents were ranked in descending order according to the total weights. Figure 4 shows the comparison of using various approaches for the same dataset and query 46. The response of using mere normalized average TF \* IDF values is low. But instead if the combined value of normalized average TF \* IDF and correlation delta is used, and then higher precision can be obtained for longer recall intervals.

Using document delta values given by equation (4), produces better results as expected than mere normalized average TF \* IDF. Alternately, vector space representation of the reference domain vs. document matrix can be used which produces comparable but slightly lower performance than using document delta approach. Finally, Singular Value Decomposition (SVD) [12] can be applied with LSI to the formed vector space in order to obtain the hidden latent structure in the dataset. The performance of precision-recall curve in response to query number 46 was not better than the VSM or the document delta approach. The reason of poor performance by SVD technique is currently being studied. All the four responses have been shown in Figure 4. Figure 5 shows the mean average precision comparison of using different approaches for first 10 MEDLINE queries.

In the experiments, mainly 2 other approaches were compared with the document delta method. In the first approach, traditional term vs. document matrix was formed in the Vector Space and mean average precision was observed. In the second approach, LSI technique was applied to term vs. document matrix using SVD and mean average precision was recorded. Both VSM and LSI results were obtained using Text to Matrix Generator software [13]. In the document delta approach, document delta was calculated for each relevant document. All three approaches were compared. Document delta approach

produces certainly better results than traditional Vector Space Model of term vs. document matrix. Also document delta approach mean average precision values are comparable to using LSI technique.



**Figure 5** Mean average precision of first 10 MEDLINE queries

The proposed document delta approach and VSM approach produce satisfactory, context-aware as well as intuitive retrieval results. The models have theoretical basis and are able to exploit dictionary references for learning different meanings as well as disambiguation.

Consider 100 documents with an average of 1000 sentences per document. If each sentence on an average consists of 5 words, TF \* IDF matrix would consist of 500000 rows and 100 columns. Less frequently appearing terms can be filtered to reduce the number of rows. Even with stemming and filtering, the resulting matrix would be large to solve. The success in the reduction of LSI matrix also depends on the sparsity of the matrix i.e. number of non-zero elements [14]. The EBOTS model uses pseudo-representation of 1000 rows and 100 columns. Considerable matrix size reduction can be achieved without a compromise with acquired knowledge using the proposed EBOTS approach.

## 5. Conclusion and Future Work

In this paper, a novel approach is presented to highlight contextual knowledge representation of text data. The proposed formal models are also effective in determining relevance to a specific context. This relevance in turn can be treated as weight for Information Retrieval process. The impact of using LSI approach with correlation delta factors is currently being studied.

The use of lexical resources such as a thesaurus and/or dictionary allows the representation of different synonyms and domain terms. Cross-language dictionary mapping will allow multi-lingual context representation. Experiments were conducted using the EBOTS system to include context-aware knowledge management. More intuitive results can be obtained using context-based retrieval over traditional retrieval systems. Further improvements can be made to the context-based retrieval system using context-based reasoning methodologies.

Context-based word sense disambiguation will yield a better understanding of contextual changes and also of the scope of the context.

## 6. References

- [1] Lyman, P., Varian, H., Charles, P., Good, N., Jordan, L. L., Pal, J.: How much Information? 2003. SIMS Lab at University of Berkeley on the web at <http://www.sims.berkeley.edu/research/projects/how-much-info-2003/> Jan18, 05
- [2] Bagnall, B., : borges2003 blog. On the web at <http://borges2003.typepad.com/borges2003/2004/12/yes-that-would.html> Jan20, 2005
- [3] Aitchison, J.: Words in the mind: an introduction to mental lexicon. 2<sup>nd</sup> Edition, Oxford and New York: Basil Blackwell (1994) pg. 39
- [4] Lesk, M. E., 1986, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone", In proceedings of Special Interest Group for Documentation Conference , Toronto, Canada, pp. 24-26.
- [5] Raghavan, V.V. , Jung, G.S., Bollman, P., A critical investigation of recall and precision as measures of retrieval system performance, ACM Transactions on Information Systems, 7(3):205-229, 1989
- [6] Fuhr, N., XIRQL: An Extension of XQL for Information Retrieval, online at <http://www.haifa.il.ibm.com/sigir00-xml/final-papers/KaiGross/sigir00.html> April 09, 2005
- [7] Roy, P., Mohania, M., Shree Raman, Context-Oriented Structured and Unstructured Information Integration Using SCORE, Technical Report, IBM India Research Lab, 2004
- [8] WordNet, lexical resource system developed by Princeton University, Cognitive Science Laboratory, online at <http://wordnet.princeton.edu/>
- [9] Porter, M.F., An algorithm for suffix stripping. Program, 1980. 14(3): p. 130-137.
- [10] Classic3 dataset for experiments was used from <ftp://ftp.cs.cornell.edu/pub/smart/>
- [11] Ando, R.K., Latent Semantic Space: Iterative Scaling Improves Precision of Inter-document Similarity Measurement. In Proceedings of the SIGIR, (Athens, Greece, 2000), 216--223.
- [12] Dumais, S. : Enhancing Performance in Latest Semantic Indexing (LSI) Retrieval – DRAFT (1992)
- [13] Zeimpekis, D., Gallopoulos, E., Text to Matrix Generator, online at <http://scgroup.hpclab.ceid.upatras.gr/scgroup/Projects/TMG/>
- [14] Dumais, S., T., Platt, J., Heckerman, D., and Sahami, M., Inductive Learning Algorithms and Representations for Text Categorization. In Proceedings of the 7<sup>th</sup> International Conference on Information and Knowledge Management. (Bethesda, Maryland, 1998)