“An Analysis on Concept Based Retrieval and Interpretation for Large Datasets”

Shiv Kumar Goyal¹, Jyoti T tahalramani², Vinita Ardeja³

¹ Deputy Head of the Department, Master of Computer Application
²Research Scholars, Master of Computer Application
³Research Scholars, Master of Computer Application

Vivekanand Educational Society’s Institute Of Technology, Chembur
Mumbai, Maharashtra

Shivkumar.goel@ves.ac.in¹
tahalramanjyoti@gmail.com²
ardejavinita@gmail.com³

Abstract:
In present paper we done analysis on Information retrieval systems traditionally rely on textual keywords to index and retrieve documents. Keyword-based retrieval may return inaccurate and incomplete results when different keywords are used to describe the same concept in the documents and in the queries. Concept-based retrieval methods have attempted to tackle these difficulties by using manually built thesauri, by relying on term co-occurrence data, or by extracting latent word relationships and concepts from a corpus. In this article we introduce a new concept-based retrieval approach based on Explicit Semantic Analysis (ESA), a recently proposed method that augments keyword-based text representation with concept-based features, automatically extracted from massive human knowledge repositories such as Wikipedia. Our approach generates new text features automatically, and we have found that high-quality feature selection becomes crucial in this setting to make the retrieval more focused.

Key Words: Concept-based retrieval, explicit semantic analysis, feature selection, semantic search.

1. INTRODUCTION

Information Retrieval (IR) systems aim at providing the most relevant documents to a user's query. Initial IR methodology was based on keywords manually assigned to documents, and on complicated Boolean queries. Documents were indexed by automatically considering all terms in them as independent keywords, in what is known as the Bag-of-Words (BOW) representation, and query formatting was simplified to a short natural language formulation. The keywords chosen by users were often different from those used by the authors of the relevant documents, lowering the systems' recall rates. In other cases, the contextual differences between ambiguous keywords were overlooked by the BOW approach, reducing the precision of the results. These two problems are commonly referred to as synonymy and polysemy, respectively.

Concept-based information retrieval is an alternative IR approach that aims to tackle these problems differently. Concept-based IR represents both documents and queries using semantic concepts, instead of keywords, and performs retrieval in that concept space. This approach holds the promise that representing documents and queries (or augmenting their BOW representation) using high-level concepts will result in a retrieval model that is less dependent on the specific terms used. Such a model could yield matches even when the same notion is described by different terms in the query and target documents, thus alleviating the synonymy problem and increasing recall. Similarly, if the correct concepts are chosen for ambiguous words
appearing in the query and in the documents, non-relevant documents that were retrieved with the BOW approach could be eliminated from the results, thus alleviating the polysemy problem and increasing precision.

In this article we propose a novel concept-based IR approach that meets all of the requirements, using Explicit Semantic Analysis (ESA) to augment the standard BOW representation. The concepts used are taken from a very comprehensive, human-defined ontology of explicit concepts. Text analysis methods are used to automatically and efficiently extract these concepts and represent any document or query text using them. Finally, the proposed system builds upon existing IR methodology and augments BOW representation with concepts in indexing and retrieval, using standard data structures and ranking methods.

Our main contributions in this work are threefold: a framework for using the ESA representation method in information retrieval, a method for integrating feature selection into the concept-based IR task, and three selection methods that are based on common AI methods and shown beneficial for the task at hand.

2. BACKGROUND

Explicit Semantic Analysis, or ESA, is a recently proposed method for semantic representation of general-domain natural language texts. ESA represents meaning in a high-dimensional space of concepts, automatically derived from large-scale human-built repositories such as Wikipedia.

A concept is generated from a single Wikipedia article, and is represented as a vector of words that occur in this article weighted by their tf.idf score. Once these concept vectors are generated, an inverted index is created to map back from each word to the concepts it is associated with. Thus, each word appearing in the Wikipedia corpus can be seen as triggering each of the concepts it points to in the inverted index, with the attached weight representing the degree of association between that word and the concept.

Figure 1: Generation of an ESA model.

The articles and words in them are processed to build a weighted inverted index, representing each word as a vector in the space of all Wikipedia concepts. As an example, these are the top ten concepts triggered by the word “investor”:

1. INVESTMENT;
2. ANGEL INVESTOR;
3. STOCK TRADER;
4. MUTUAL FUND;
5. MARGIN (FINANCE);
6. MODERN PORTFOLIO THEORY;
7. EQUITY INVESTMENT;
8. EXCHANGE-TRADED FUND;
9. HEDGE FUND;
10. PONZI SCHEME.

We believe that the use of a knowledge repository as large and diverse as Wikipedia creates a powerful concept ontology, well suited for concept-based IR.

3. ESA-BASED RETRIEVAL
In this section we introduce our first algorithm for concept-based IR using ESA representation. The algorithm maps documents and queries to the Wikipedia-ESA concept space, and performs indexing and retrieval in that space. We then evaluate the algorithm’s performance on TREC datasets. We show that combining concept-based relevancy of documents with that of passages in these documents performs best for ESA-based retrieval.

3.1. ESA Concept-Based Indexing

We use ESA to map each document in the corpus to a weighted vector of concepts. Like BOW vectors, concept-based vectors are also sparse, with concept weights being zero for most of the Wikipedia concepts. Indexing the entire list of related concepts for every document is not feasible. We therefore use only those concepts with the highest weights (association scores). In a sorted representation of the vector, this subset of concepts is simply its prefix. Long documents are more difficult to map in full into the ESA concept space. Previous research using BOW representation has shown that breaking long documents into shorter passages can improve document retrieval with the ranking of passages viewed as evidence to the relevance of their source documents. Furthermore, fixed-length passages yield better results than passages based on syntactic or semantic segmentation. In our approach, each passage is indexed and may be retrieved as a stand-alone unit of information. Thus, a passage is ranked separately as an independent indicator of its original document’s relevance. We now have, for any document to be indexed, a set of passages and a concept vector representation for each. We index these concepts in a standard IR inverted index, using the concepts’ unique identifiers as tokens. The score associated with each concept in the vector is used as the token weight, equivalent to term frequency in standard text indexing. The pseudo code for the preceding indexing algorithm is described in Figure 2.

![Figure 2: ESA-based indexing in an inverted index](image)

3.2. ESA-Based Retrieval Algorithm

Upon receiving a query, our algorithm first converts it to an ESA concept vector. Having indexed full documents and passages, we now have to choose how these two types of evidence are to be combined for ranking. Following we retrieve both sets of results and sum each document’s full score with the score of the best performing passage in it. The documents are then sorted by this combined score and the top scoring documents are output, as described in Figure 3. The retrieval algorithm has a single parameter $s$ controlling the cut-off (as described in the previous section).
The retrieval algorithm has a single parameter $s$ controlling the cut-off (as described in the previous section) of the query concept vector. The value for $s$ may be chosen to be the same as that in the indexing process, but not necessarily. Indexing the entire corpus with large cut-off values would incur significant storage and computation costs, and is therefore not feasible. The query representation, on the other hand, being derived from a much shorter text fragment and incurring no such costs, could benefit from a finer representation, using a higher value for $s$.

3.3 Empirical Evaluation

In order to evaluate the usefulness of ESA concept-based retrieval, we carried out a set of experiments.

3.3.1 Implementation

We used Xapian3, an open source probabilistic IR library, as the basis for our experimental platform. Document keywords and concepts were indexed in a Xapian inverted index. Most of the experiments used the TREC-8 Adcock and the TREC Robust 2004 datasets. We used only the short queries in TREC topics, since these short (1-3 words) queries better represent common real-life queries and since short texts stand best. We use the Mean Average Precision (MAP) evaluation measure, commonly used by TREC participants who combine precision and recall while assigning higher importance to the higher-ranking relevant documents.

Documents and passages were stemmed, stopped and indexed by their BOW representation, to serve as the keyword baseline index. Then, ESA-based representations were created and indexed separately as described in Figure 3.1. Passages were set to be fixed-size overlapping segments, shown to be most effective by with passage size set to 50 words.

3.3.2 Results

In this paper the performance (MAP) of our ESA-based retrieval algorithm for various parameter values. To assess the impact of the concept vector truncation, we measured performance for varying values of $s$ (the ESA vector cut-off level) in the query vector. In addition, to validate the added value of combining documents and passages scores, we compared performance of the combined score to that of documents and passages alone.

As Figure 4 clearly shows, passage context outperforms document context significantly, but the best results are achieved when both are combined, an outcome that is consistent with previous IR endings for BOW representations. We will be using the
combined documents passages scoring from here onwards.

Results for increasing values of s indicate that merely adding lower-ranking concepts in the ESA vector does not improve retrieval. Not only does the precision-oriented MAP score decrease as concepts are added, but the absolute recall (measured in the top 1000 retrieved documents) decreases as well. This finding suggests that some of the generated concepts may be detrimental, and that successful application of ESA to IR may require further selection of the concepts initially generated for the query.

"ANCIENT ARTIFACTS FOUND. Divers have recovered artefacts lying underwater for more than 2,000 years in the wreck of a Roman ship that sank in the Gulf of Baratti, 12 miles of the island of Elba, newspapers reported Saturday."

The top 10 concepts generated for this document were:

- Scuba diving
- Wreck diving
- RMS Titanic
- USS Hoel (DD-533)
- Shipwreck
- Underwater archaeology
- USS Maine (ACR-1)
- Maritime archaeology
- Tomb Raider II
- USS Meade (DD-602)

Whereas the query's top 10 concepts were:

- Shipwreck
- Treasure
- Maritime archaeology
- Marine salvage
- History of the British Virgin Islands
- Wrecking (shipwreck)
- Key West, Florida
- Flotsam and jetsam
- Wreck diving
- Spanish treasure fleet

With 3 matches in the top-10 concepts (and more in lower positions), the ESA-based method was capable of retrieving this relevant document as its third ranked result, despite the fact that not one of the query terms appears in the document's text.

Our observation is that since the ESA classifier is created from a noisy unstructured information source, and one that is different from the target corpus, the initial concept

---

3.3.3 Qualitative Analysis

The results show that ESA-based retrieval can indeed, as expected, identify relevant documents even when these do not include query terms or their simple synonyms. Let us consider TREC query 411 ("salvaging shipwreck treasure"). The following short relevant document was retrieved by the ESA-based method but not by the BOW baseline:

"ANCIENT ARTIFACTS FOUND. Divers have recovered artefacts lying underwater for more than 2,000 years in the wreck of a Roman ship that sank in the Gulf of Baratti, 12 miles of the island of Elba, newspapers reported Saturday."

The top 10 concepts generated for this document were:

- Scuba diving
- Wreck diving
- RMS Titanic
- USS Hoel (DD-533)
- Shipwreck
- Underwater archaeology
- USS Maine (ACR-1)
- Maritime archaeology
- Tomb Raider II
- USS Meade (DD-602)

Whereas the query's top 10 concepts were:

- Shipwreck
- Treasure
- Maritime archaeology
- Marine salvage
- History of the British Virgin Islands
- Wrecking (shipwreck)
- Key West, Florida
- Flotsam and jetsam
- Wreck diving
- Spanish treasure fleet

With 3 matches in the top-10 concepts (and more in lower positions), the ESA-based method was capable of retrieving this relevant document as its third ranked result, despite the fact that not one of the query terms appears in the document's text.

Our observation is that since the ESA classifier is created from a noisy unstructured information source, and one that is different from the target corpus, the initial concept
vector might carry noise and ambiguities. To counter such problems, we hypothesized that the concept vector should first be tuned to better fit the corpus it is querying. This is similar to the idea that a corpus-based similarity thesaurus is better than a general purpose one.

An ESA vector has two candidates for such tuning: the subset of concepts and the weights assigned to them. To check whether tuning should be performed for both of them, we ran the same tests as before, but with all query concept weights set to a uniform value. We found that this change hardly made any difference in performance, and this conclusion was also verified in similar tests in later experiments. Thus, we conclude that tuning the original concepts is useful only when altering the set of concepts to be used.

4. SELECTIVE ESA-BASED RETRIEVAL

4.1 Feature Selection using Pseudo-Relevance Feedback

When ESA was applied to the text categorization task, it was vulnerable to some problems. Nevertheless, the researchers overcame these problems by employing aggressive feature selection (FS). FS methods use labelled training examples to evaluate the utility of candidate features. In contrast, the IR task inherently lacks any labelled training data; hence applying FS to information retrieval will require finding an alternative method of evaluating the utility of features (concepts in our case).

For this purpose, we note that there exists also a supervised version of IR, called relevance feedback, where the user provides relevance judgements on an initial set of retrieved results. This feedback is then used to reformulate the query and retrieve an improved set of results. Relevance feedback can be extended to the unsupervised case, by assuming that the top ranked results (documents or passages) in the initial retrieved results are relevant. This method is commonly referred to as pseudo-relevance feedback (PRF). Inspired by PRF, we decided to use the results of keyword-based retrieval as a source for evaluation in our FS process. Our updated retrieval method will thus become two-phased, first performing keyword-based retrieval, then using its results to tune the query concepts and perform concept-based retrieval.

Next, we had to decide which subsets of the results are to be used. Most of the work on PRF used the top ranked documents or passages as pseudo-relevant documents (or positive examples). Some researchers chose to include also pseudo-non-relevant documents (or negative examples), by using the bottom-ranked documents. We chose to use both positive and negative examples, as the initial query representation includes irrelevant concepts to be removed.

4.2 Selective ESA-Based Retrieval Algorithm

Now that we have decided on a framework for evaluating features, let us describe the integration of FS into the general ESA-based retrieval algorithm.
First, as in the non-selective algorithm, the textual query \( q \) is represented by an ESA concept vector \( F_q \). Then, the first \( n \) results ranked by keyword-based retrieval for \( q \) are fetched. The top \( k \) of these \( (k << n) \) are tagged as pseudo-relevant, or positive examples, and the bottom \( k \) are tagged as pseudo-non-relevant, or negative examples. Feature selection is then applied to these examples in order to select the best performing concepts in \( F_q \), resulting in a modified ESA vector \( F_0 q \). Finally, concept-based retrieval is performed using \( F_0 q \) and results are returned. The entire process is illustrated in Figure 6.

**Figure 5**: Selective ESA-based retrieval

**Figure 6**: The PRF-based feature selection process

### 4.3.1 Methodology

We continue using the experimental framework described in Section 3.3.1, and evaluate each suggested selection method with various system parameter settings.

### 4.3.2 IG Method Results

The IG method has two primary parameters: the number of pseudo-relevant examples \( (k) \) and the selection level \( (\mu) \). Figure 11 shows retrieval performance (averaged over all queries in each dataset) as a function of \( \mu \) for several values of \( k \), compared with a baseline that performs no FS at all. Both datasets show similar behaviour, with FS performance consistently improving as selection level increases, peaking at \( \mu = 20\% \) (which implies retaining 10 out of the initial 50 features). More aggressive selection is already damaging, probably as the result of removing useful features along with non-relevant ones.
4.3.3 IIG Method Results

The IIG method requires only one parameter to be set, the size of the positive/negative example set \( (k) \). In addition, the algorithm may be run in forward-selection or in backward-elimination mode. Figure 12 shows retrieval results for different values of \( k \) in both modes, compared with results of the initial baseline query.

![Figure 11: Concept-based performance as a function of a fraction of the concepts selected (\( \mu \)), IG method](image)

![Figure 12: Concept-based performance as a function of the number of pseudo-relevant examples (\( k \)), IIG method](image)

4.3.4 RV Method Results

The RV method, like IG, requires setting two parameters, \( k \) and \( \mu \). Like the graphs in the previous sections, the graphs in Figure 13 show the impact of these parameters on the system's performance. But whereas with the IG method the query reverts to the original query at \( \mu = 100\% \), this is not the case with the RV method. Even without any selection, the query changes as a result of adding the features generated from the positive example documents and of the reweighting step.
4.3.5 Parameter Tuning through Training

All selection methods shown in this section rely on one or two system parameters, whose values may have a significant impact on system performance. These parameters can be tuned if a set of queries is provided with relevance judgments on result documents. We operated the system on TREC-7 queries with the three proposed FS methods and varied the parameter value ranges. The resulting best performance values obtained were: for IG FS \(hk = 10; \mu = 30\%\), for IIGFS forward-selection; \(k = 10\) and for RV FS \(hk = 35; \mu = 20\%\).

4.4 Analysis

The IG method exhibits good peak behaviour, but it seems to be highly sensitive to the chosen selection level \(\mu\). Tuning the system parameters using training data, if available, may significantly alleviate this problem. The IIG forward-selection method appears to perform better than backward-elimination. This method requires tuning only a single parameter. It would therefore be the preferred choice when no training data is available. Its performance, though, is slightly lower than IG, and it is still quite sensitive to the \(k\) parameter value.

The RV method performs slightly worse than IG for small example setsizes, probably due to its overdependence on the quality of these examples. For larger example sets (in our case, \(k > 15\)), it performs comparably to the IG method.

Related Work

Representing texts using concepts that are words, or explicit syntactic/semantic classes, has the benefit of producing concepts that are human-readable, easy to analyze and reason about, and can be displayed to a user of such a system. ESA concepts, too, are based on human defined natural concepts, as the example concept names throughout this paper show. Yet concepts may also be defined using latent semantics, with possibly broader concept coverage. By analyzing the latent relationships between terms in the target corpus, methods such as Latent Semantic Indexing (LSI) can project the term space to a reduced-dimensions concept space, shared by documents and queries, and thus be applied successfully to the IR task. More recent dimensionality reduction methods applied to IR have included TopicModels approaches such as Latent Dirichlet Allocation and the Pachinko Allocation Model.

Conclusion
Concept-based IR using ESA makes use of concepts that encompass human world knowledge, encoded into resources such as Wikipedia (from which an ESA model is generated), and that allow intuitive reasoning and analysis. Feature selection is applied to the query concepts to optimize the representation and remove noise and ambiguity. The results obtained by our proposed system (Morag) are significantly better than the baselines used, including those of top performing systems in TREC-8. Analysis of the results shows that improving the performance of the FS component is possible and will directly lead to even better results.

**Bibliography**


