Contents

[Objective 1](#_Toc373673412)

[Theoretical Phase 1](#_Toc373673413)

[Information Retrieval 1](#_Toc373673414)

[Pseudo Relevance Feedback 2](#_Toc373673415)

[Smoothing 2](#_Toc373673416)

[Statistical Methods 2](#_Toc373673417)

[Probabilistic Language Models 2](#_Toc373673418)

[Language Models 2](#_Toc373673419)

[An improvement: KL Divergence Model 3](#_Toc373673420)

[Document Model 3](#_Toc373673421)

[Query Model 3](#_Toc373673422)

[Maximum Likelihood 4](#_Toc373673423)

[Document Model 4](#_Toc373673424)

[Conceptual Query Models 4](#_Toc373673425)

[Cluster Based Retrieval 5](#_Toc373673426)

[Semantic Models 6](#_Toc373673427)

[Implementation Phase 7](#_Toc373673428)

[Language conceptual model 7](#_Toc373673429)

[Drawbacks of Language conceptual model method- 8](#_Toc373673430)

[WFIDF and TFIDF methods: 8](#_Toc373673431)

[Drawbacks of WFIDF / TFIDF Methods: 9](#_Toc373673432)

[Semantic Model 9](#_Toc373673433)

[Implementation 10](#_Toc373673434)

[Drawbacks of Semantic Model 10](#_Toc373673435)

[Results and conclusion 11](#_Toc373673436)

[References 12](#_Toc373673437)

# Objective

With given collection of documents and concepts for each document, this project analyses different language models- conceptual model, WFIDF / TFIDF Methods, semantic language model based on already established theoretical papers and formula. We study on the drawbacks and limitations of each method and offer optimal solutions. In this project, we aim to find what is the probability of occurrence of a concept for a given query and hence extracting important relevant concepts for each document.

# Theoretical Phase

The basic understanding of the work done till present in the field of Information Retrieval contributed to the major part of the Study Oriented Project. We concentrated upon the following aspects:

* Prevalent Models used in Language Modeling.
* Algorithms used by search engines.
* Limitations and advantages of one model over the other.
* Performance and correctness

The following papers were analyzed to gain the latest information in the field of information retrieval:

* Conceptual language models for domain-specific retrieval

Edgar Meij, Dolf Trieschnig, Maarten de Rijke, Wessel Kraaij

* Cluster-Based Retrieval Using Language Models

Glimpse of the findings in theoretical phase is documented under the following sections. Important terminology and techniques are explained.

## Information Retrieval

Queries are formal statements of [information needs](http://en.wikipedia.org/wiki/Information_need). In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of [relevancy](http://en.wikipedia.org/wiki/Relevance). Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

## Pseudo Relevance Feedback

It automates the manual part of relevance feedback, so that the user gets improved retrieval performance without an extended interaction. The method is to do normal retrieval to find an initial set of most relevant documents, to then assume that the top "k" ranked documents are relevant, and finally to do relevance feedback as before under this assumption.

## Smoothing

Smoothing is applied both to avoid zero-frequency problems occurring with a MLE approach and to account for general and document-specific language use.  
A common approach is to generate a maximum-likelihood model for the entire collection and linearly interpolate the collection model with a ML model for each document to create a smoothed document model.  


## Statistical Methods

### Probabilistic Language Models

To compute the probability of a sentence or sequence of words:  
P(W) = P(w1,w2,w3,w4,w5….. wn)

Probability of an upcoming word: P(w5|w1,w2,w3,w4)

### Language Models

Chain Rule:

Description: \begin{displaymath}
P(t_1t_2t_3t_4) = P(t_1)P(t_2\vert t_1)P(t_3\vert t_1t_2)P(t_4\vert t_1t_2t_3)
\end{displaymath}

Unigram Language Model

Description: \begin{displaymath}
P_{uni}(t_1t_2t_3t_4) = P(t_1)P(t_2)P(t_3)P(t_4)
\end{displaymath}

Bigram Language Model

Description: \begin{displaymath}
P_{bi}(t_1t_2t_3t_4) = P(t_1)P(t_2\vert t_1)P(t_3\vert t_2)P(t_4\vert t_3)
\end{displaymath}

Multinomial Unigram Language Model   


With limited training data, a more constrained model tends to perform better. Unigram models are more efficient to estimate and apply than higher-order models.

### An improvement: KL Divergence Model

Using KL-divergence, documents are scored by measuring the divergence between a query model and each document model.  

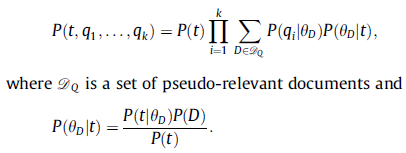

**Query specific constant**

## Document Model

Each document is considered to be a mixture of a document-specific model and a more general background model. Thus, each document model is estimated as the MLE of each term in the document P(t|D), linearly interpolated with a background language model P(t), which in turn is calculated as the likelihood of observing t in a sufficiently large corpus, such as the document collection. 

## Query Model

Query Models are approximation of Relevance Models which rely heavily on relevance feedback. Since there might be very less or no training data, two methods are presented for constructing models from the queries and a set of pseudo-relevant documents, using different independence assumptions.

* Method 1 : Probability of observing t after observing q:
* Method 2 : The query terms are independent of each other, but keep their dependence on t : 

## Maximum Likelihood

When the query model is generated using the empirical, maximum-likelihood estimate (MLE) on the original query, i.e.,



It can be shown that documents are ranked in the same order as using the query likelihood model.

## Document Model

Each document is considered to be a mixture of a document-specific model and a more general background model. Thus, each document model is estimated as the MLE of each term in the document P(t|D), linearly interpolated with a background language model P(t), which in turn is calculated as the likelihood of observing t in a sufficiently large corpus, such as the document collection. 

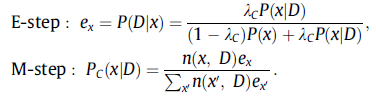
## Conceptual Query Models

**The goal** is to utilize the knowledge that is encapsulated in a concept language to enhance the estimation of the query model.

  
Enhances the estimation of extended part of query model.

* The second component—the conceptual query model P(c|Q)—is a distribution over concepts specific to the query. In some settings, concepts are provided with a query or as part of a query. If this is not the case, however, we may leverage the document annotations to approximate this step. The estimation of concepts relevant to a query in a standard language modeling manner, by determining which concepts are most likely given documents relevant to the query. For this, we use *pseudo-relevance* methods.  
  
* The first term defines the parameters of a generative language model for each concept : a generative concept model. While the concepts may be directly usable for retrieving documents, we associate each concept with a weighed set of most characteristic terms using a multinomial unigram model. To this end we consider the documents that are annotated using c as bridges between the concept and terms, by representing concepts as multinomial distributions over terms, P(t|c). The strength of association between a concept c and a term t by determining the probability of observing t given c:   
  
* It may be assumed that t and c are independent and identical samples given a document D in (or with) which they occur. So, the probability of observing both t and c is :



* Setting up the parameters:  
  
* The following EM algorithm until the estimates do not change significantly anymore:  
  

## Cluster Based Retrieval

* The [cluster hypothesis](http://en.wikipedia.org/w/index.php?title=Cluster_hypothesis&action=edit&redlink=1) asserts that two documents that are similar to each other have a high likelihood of being relevant to the same information need. Two documents in the same cluster should both be relevant to the same request.  
  Cluster-based retrieval requires that documents be first organized into clusters.
* To cluster documents one must establish
* A pairwise measure of document similarity (or distance).
* Then choose a clustering algorithm to group documents based on their similarity(or distance).
* cosine measure for document similarity is opted.

Static clustering methods

* Clustering, independent of the user’s query.
* Clusters are retrieved based on how well their centroids match the query.
* Partitioning methods are used.

A three-pass K-means algorithm-

* K is an input : the desired number of clusters.
* In the first pass, the algorithm takes the first K documents each as the centroid of a unique cluster.
* Each of the remaining documents is then compared to these centroids and assigned to the cluster with the closest centroid.
* In the following passes, the cluster centroids are recomputed based on clusters formed in the previous pass and the cluster-membership of documents are re-evaluated based on these new centroids.
* The running time for each pass is linear in the total number (N) of documents to be clustered, i.e. *O(K*⋅*N)*.

**CBDM Model:**  
 It uses clusters to smooth documents.  


**TDM Model:**  
The CBDM model can also be viewed as a mixture model of three sources: the document, the cluster/topic the document belongs to, and the collection. A relevant document is assumed being generated by this mixture model.   
****

## Semantic Models

* One of the first attempts at automatically relating concepts with text was introduced in the 1980s. Giger (1988) incorporated a mapping between concepts from a thesaurus and words as they appear in the collection. The main motivation was to move beyond text-based retrieval and bridge the semantic gap between the user and the information retrieval system, a motivation closely related to ours.

# Implementation Phase

## Language conceptual model

Input:

Dataset directory with multiple folders. Each folder has 2 files-

1) Concepts\_unformatted: Each line represents concepts for that document.

2) Snip\_final: Each line represents one document.

Ouput:

P(c/q) = Given a query, what is the probability of occurrence of a concept for that collection

Formulae used for calculations:

C= Collection of documents

c=concepts of the collection

q=query

t=term (each term seen in the collection of documents)

D=Document of the collection

T’= each term present in the document

D’=each document of the collection

N(t,Q)= frequency of term in the query

### Drawbacks of Language conceptual model method-

1) If a document does not contain even one of the query terms, then also p(q/d)=0 which makes p(t/q) and hence p(c/q)=0 for that particular document.

2) p(c/t) does not have any semantic relation, it is just calculated as p(c/d)\*p(t/d). So even if they are unrelated, we get some value based on their frequency in the document.

3) If a term is present in more documents but it’s frequency is less and vice versa, then based on parameter value, p(t/d) holds unduly advantages and disadvantages.

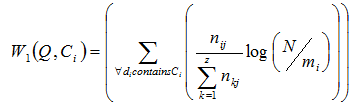
## WFIDF and TFIDF methods:

**WF:**  
The following is the formula to calculate the weight of concept with respect to query i.e P(ci/Q)



where is a number that the number of times the concept  occurs in the document , is total number of distinct terms in the document . is the total number of terms in the document . is the total number of documents considered. is the total number of documents in which the concept  appears. The weight is divided by the number of documents considered i.e. . The concepts, for which the weights are greater than the threshold , are considered to be relevant to the query.

**TFIDF:**



### Drawbacks of WFIDF / TFIDF Methods:

W(Q,C) does not have any semantic relation. So even if they are unrelated, we get some value based on their frequency in the document. Or in the case they are Synonyms/ Hyponyms/ Hypernyms of each other we won’t see the corresponding effect.

## Semantic Model

Input:  
Dataset directory with multiple folders. Each folder has 2 files-

1) Concepts\_unformatted: Each line represents concepts for that document.

2) Snip\_final: Each line represents one document.

3) Wordnet Library

Output:

P(c/q) = Given a query, what is its similarity measure with a given concept for that collection.

Similarity Measure

If we come across two words, 'car' and 'bicycle', we know they are related as both are means of transport. Also, we easily observe that 'bicycle' is more related to 'car' than 'fork' is. Similarity measure assigns quantitative value to this measure.

Given any formula rel(c1,c2) for semantic relatedness between two concepts c1 and c2,the relatedness rel between two words and can be calculated as   


where rel(c1, c2) can be on the basis of one of the following measures of similarity :

* Synonyms
* Antonyms
* Hypernamy
* Hyponamy

We are using **Resnick’s** information based approach which lies upon the intuition that one criterion of similarity between two concepts is “the extent to which they share information in common”. It can be modeled mathematically as :  
  
let p(t) be the probability of encountering an instance of a term t, then information content in c will be   
IC(t) = - log p(t)   
We can define the semantic similarity of a pair of terms as:  
  
sim (w1,w2) = - ( log p )\*( rel (w1,w2) )  
  
w1 and w2 are query and concept terms respectively.

### Implementation

Based on the theoretical analysis of conceptual language model and paper given, simulations on Java were run with some modifications.

Steps-

* Wordnet Library was synced with Java using JWNL.
* The implementation of Similarity Measure taken from Wordnet :: Similarity which is implemented in Perl, with the help of Inline Java for translation.
* Semantic relatedness between Query and Concept was measured using the above explained equations and method.
* The concepts were shown related to the query if the measure of Semantic relatedness exceeds a threshold value which was fixed to be 0.5.

### Drawbacks of Semantic Model

* Some Noun Phrases which might have been commonly used in the document might not be present in the dictionary, and hence always value less than threshold, irrespective of their high frequency of appearance.
* Semantic relatedness is checked only with individual terms but not with entire query/ concept phrase.

# Results and conclusion

1) A new formula should be devised to calculate p(q/d) for a document, giving up the strictness of mandatory presence of each term.

2) Semantic relationship between concept and terms should be considered and hence, thesaurus should be incorporated in language conceptual model.

3) In semantic model, it is important to identify proper nouns which can not be incorporated in WordNet, and semantic relatedness measure for phrases instead of single terms should also be introduced.

# References

* Cluster-Based Retrieval Using Language Models.
* Meij, E., et al. Conceptual language models for domain-specific retrieval. Information Processing and Management (2009), doi:10.1016/j.ipm.2009.09.005
* Zhai, C., & Lafferty, J. (2002). Two-Stage language models for information retrieval. In *SIGIR 2002*, pp. 49-56
* Croft W. B., & Lafferty, J (eds.) (2003). *Language Modeling for Information Retrieval*. In Kluwer International Series on Information Retrieval, Volume 13, Kluwer Academic Publishers.
* Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In SIGIR ’01.
* Lafferty, J., & Zhai, C. (2003). Probabilistic relevance models based on document and query generation. In Language modeling for information retrieval Springer.
* [Budanitsky and Hirst, 2001]. EvaluatingWordNet-basedMeasures of Lexical Semantic Relatedness.
* <https://metacpan.org/pod/WordNet::Similarity>
* P.Resnik Wordnet and class based probabilities.
* Wordnet.princeton.edu
* Wikipedia