**CS6200-Information Retrieval**

**(Fall 2016)**

**Members**

Ashok Koduru

Sravanthi Kethireddy

Frenia Pinto

**Instructor**

Professor Nada Naji

**Introduction**

The project demonstrates designing and building information retrieval systems, evaluating and comparing their performance levels in terms of retrieval effectiveness. The algorithms that are implemented for retrieval are:

1. Tf-idf
2. Cosine similarity
3. Lucene
4. BM-25 algorithm

The project also includes implementation of the query expansion technique using Pseudo Relevance feedback. Using stopping and stemming on corpus, two other runs are produced.

The runs produced by the retrieval models are evaluated using:

1. MAP
2. MRR
3. P@K, K = 5 and 20
4. Precision and Recall

**Contribution of the team-members:**

Ashok:

Sravanthi:

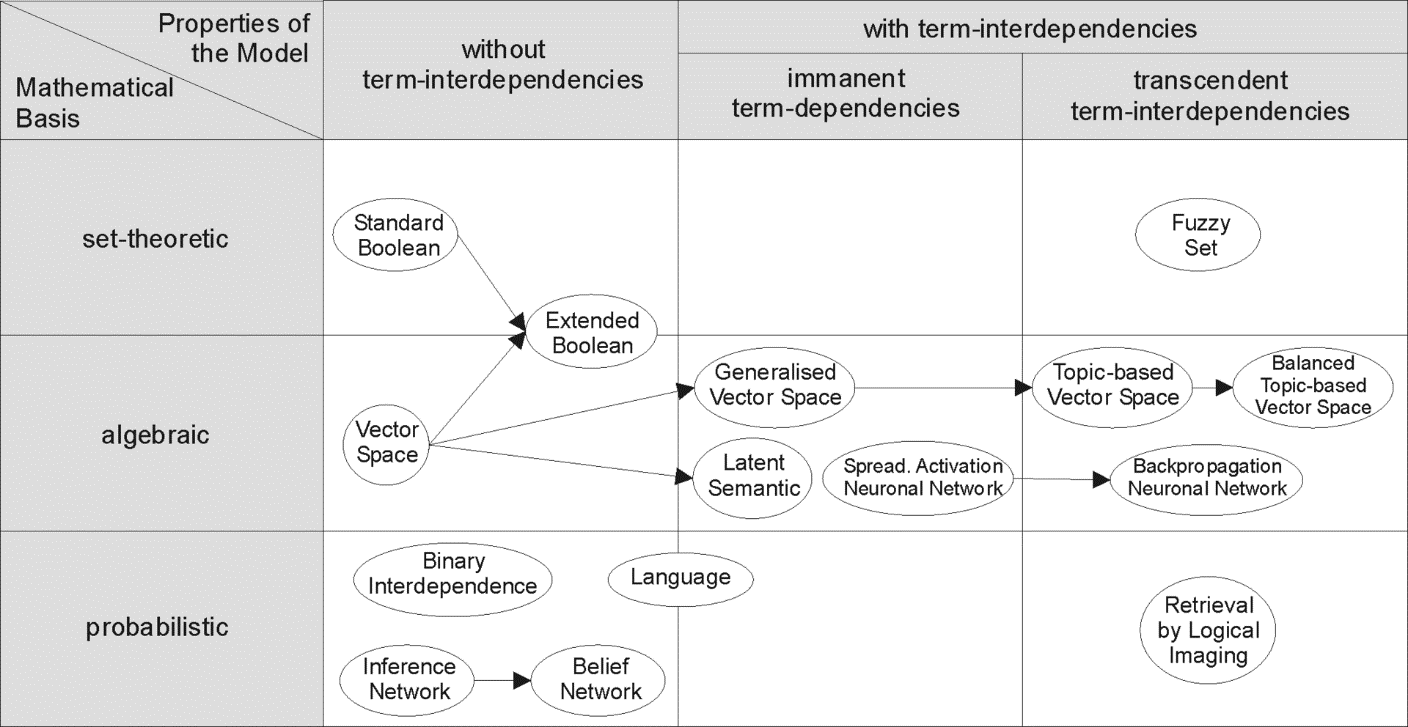
Frenia:

**Literature and Resources**

The purpose of Information Retrieval(IR) is to provide the documents which satisfy the users information need. The process of IR can be depicted in the following diagram.



There are a wide range of retrieval models which are classified as follows:



**Tf-idf retrieval model:**

The tf-idf weight is the statistical weight measure used to determine importance of the term in a corpus. The term importance is proportional to the term frequency in a document but inversely proportional to frequency of term in corpus.

Tf = (occurrence of a term in a document) / (number of terms in the document)

Idf = loge(Total number of documents / number of documents with the term t in it)

We multiplied these terms to rank the documents according to the given query.

**Cosine similarity:**

A cosine similarity is a measure of the direction-length resemblance between vectors. In order to compute cosine similarity two document vectors are needed where each unique term in the index is a vector and the value at that index is the measure of term importance in that document. This representation of set of documents as vectors is called vector space model.

Vector representation of documents and query

To retrieve the ranked documents, we utilized the vector space similarity model(VSM). The VSM score of a document d for query q is the cosine similarity of weighted query vectors V(q) and V(d).

Cosine-similarity (q, d) = V(q). V(d) / |V(q)| |V(d)|

V(q). V(d) is the dot product of the weighted vectors

|V(q)| and |V(d)| are the Euclidean norms

**Lucene:**

Lucene is an open source information retrieval software library which is developed in java. Lucene scoring is done using the TfIdf similarity model, we implemented the same and ranked the documents accordingly.

**BM25 retrieval model:**

We utilized the following formula available for BM25 to compute the ranks of the documents.

Σ log((*r* + 0.5)(*N* − *n* − *R* + *r* + 0.5))/((*n* − *r* + 0.5)(*R* − *r* + 0.5))\* ((*k*3 + 1)*q*)/((*k*3 + *q*)) \*

((*k*1 + 1)*f*)/((K + *f*))

Where:

r = number of relevant documents indexd by the term

R = total number of relevant documents

N = number of documents in the collection

n = number of documents in the collection indexed by the term

k1, k3 are constants

q = term frequency in the query

f = term frequency in the document

K = k1((1-b)+b\*(document length/ average document length))

For implementation of the query expansion, ‘Pseudo Relevance Feedback’ approach is being used. The expansion terms generated by pseudo-relevance feedback will depend on the whole query, since they are extracted from documents ranked highly for that query, but the quality of the expansion will be determined by how many of the top-ranked documents in the initial ranking are in fact relevant. The derivational/inflectional variants, thesauri, ontologies are used to generate language-specific terms.

We have used this approach because derivational/inflectional variants are used to add the variants (parts of speech) terms to the query which may change the entire meaning of the query terms entered by the user. The thesauri and ontology adds synonyms to the query terms. This approach may not allow the user to find the exact document he is looking for. Also, pseudo relevance feedback is most effective and widely used.

**Implementation and Discussion**

**Results**

**Conclusions and Outlook**

**Bibliography**

1. <http://wwwhome.cs.utwente.nl/~hiemstra/papers/IRModelsTutorial-draft.pdf>
2. <https://en.wikipedia.org/wiki/Information_retrieval#Model_types>
3. https://lucene.apache.org/core/4\_6\_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html