Modeling using Term Dependencies and Term Weighting in Information Retrieval Systems

Doaa Mabrouk, Sherine Rady ,Nagwa Badr and M.E.Khalifa

Information System Department, Faculty of Computer and Information sciences
Ain Shams University, Cairo, Egypt

Doaa.mabrouk19@gmail.com, srady@cis.asu.edu.eg
nagwa.badr@cis.asu.edu.eg, Esskhalifa@cis.asu.edu.eg

Abstract

The process of information retrieval has become a fact of life because of the great growth of the internet and information. In this paper, retrieval models are discussed. These models are classified according to two different mathematical modeling approaches: term dependency models and term weighting models. Some models assume term independence, while others consider this dependency relation. On the other hand, term weighting is a method that tries to index the document in an effective way. The aim of this paper is to discuss and evaluate different information retrieval models which employ both terms dependence/independence, as well as term weighting.

Keywords: *Information retrieval system, Mathematical modeling , Term dependency modeling, Term weighting modeling.*

1. Introduction

Information retrieval system (IRS) is a system that able to store, retrieve and maintain information, [18]. Information has many types such as text (include numeric and date data), audio, video and other multimedia. Information retrieval modeling is very important to help researchers in designing and implementing an actual efficient information system. Mathematical modeling can be used in several domains such as education, medical, mathematical sciences ….etc. The model of IR helps the user to predict and explain what a user will find relevant given query. Mathematical retrieval modeling is classified as classical models (Boolean and vector space models), probabilistic models (BM-25 and language models) and combining evidence models(inference network and language to rank models). In Figure 1, the classification of mathematical models in IR is shown**.**

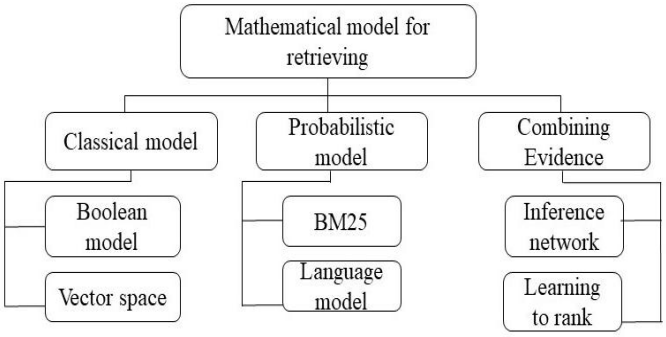

Figure 1. Mathematical Modeling Classification

1.1 Classical Models

This group contains two models: Boolean algebra and region models. In [22], these models provide exact matching. These models use logic operators (and, or, not). These operators are known as "intersection, union, difference". In Figure 2, the different Boolean model operations are shown. Their advantages are given exact match but no ranking retrieved the document. There is a difference between them which is that region model is designed to search for semi-structured data.

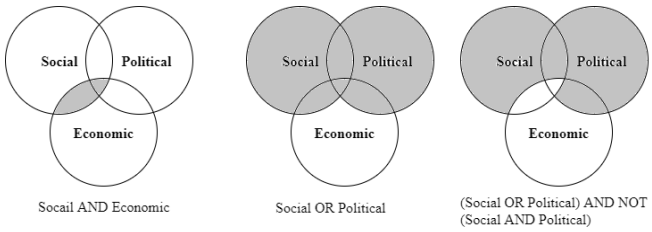

Figure 2. Boolean operation using Venn diagram [22].

The vector space model is used to calculate the similarity "distance" between document and queries through “Euclidean, Manhattan, and cosine similarity”. “If cosine angel is zero”, then the vector is orthogonal else the degree is zero. The positioning of the query in vector space model is to calculate the centroid of the relevant and irrelevant documents. Moving query towards centroid point of relevant rather than irrelevant meaning it is improving retrieval performance. Both of term weighting’s intuition and term’s independence are considered the most common disadvantages.

1.2 Probabilistic Models

Probabilistic models use conditional probability and require two conditions, [22], relevant document and long queries. The relevant document is obtained through computing the probability that contains document terms. Long queries are used to distinguish between term’s presence and term’s absence in documents. The good choice is that both relevant and non-relevant documents are available. There is a difference between probabilistic approach and knowledge based. The probabilistic approach does not require supplementary data for the data. properties and distribution while knowledge based requires some. The result of probabilistic approached is a probability while the result of knowledge based is mostly deterministic.

Two Poisson Model is developing a set of statistical rules to identify indexing term using a mixture of two Poisson, [7, 22]. This model assumes that document was created randomly as a stream of term occurrences. A number of occurrences term frequency (TF) of terms in documents can be modeled by a mixture of two Poisson distribution as follows:

$$p(x = tf) = \lambda \frac{e^{-\mu_1} - (\mu_1)^{tf}}{tf!} + (1 + \lambda) \frac{e^{-\mu_2} - (\mu_2)^{tf}}{tf!} \qquad (1)$$

Where, x is random variable for the occurrence, λ Proportion of document and μ_1, μ_2 Mean number of occurrence of term.Their advantages are that they don’t need an additional term weighting algorithm to be implemented. It is still one of the best performing term weighting algorithms.

Bayesian network models are a Probabilistic model that use a cyclic directed graph. A directed graph is acyclic if there is no directed path from A->Z. The presentation of a probability distribution as a directed graph makes it possible to analyze complex conditional independence assumption by graph theory.

The language model has been used in speech recognition, [22]. Speech recognition systems combine two probabilistic models such as acoustic model and language model. The acoustic model might produce instance for decreasing probability order. For example: "good morning", "mood morning" and so on. The language model that determines phrase is much more probable such as "good morning". It occurs more frequently in English rather than other phrases. It builds for each document. The language model, [22], assign a high probability to the word "retrieval". This is indicating that it is a good for retrieval if the query contains this word. It is helpful in that situation which requires models of similarity language or document priors.

Google page rank is a probabilistic model. It is used to determine the quality of pages, [17, 22]. The goal of this algorithm is to track some difficulties with the content-based ranking algorithms of early search engines which use text documents for webpages to retrieve the information with no explicit link relationship between them. It is called as a static ranking function. It is used in the situation that needs modeling of more of less static relations between documents. The advantages of PageRank are that it is a global measure and a query independent as well while the disadvantages are that it is very efficient to raise your own PageRank and is 'buying' a link on a page with high PageRank as well.

1.3 Combining Evidence

Learning to rank algorithm is a part of large document retrieval. It is supervised learning task. Data consists of queries and documents. It is the first trained on the training set (represent as feature vectors). It is divided into three approaches (pointwise, pairwise and listwise). Pointwise: treat ranking is as regular classification. The output of it is a class. The goal of it is to minimize the number of the wrong classification. Pairwise: It transforms ranking into pointwise classification. The goal of it is to decrease the number of pairs which are ranked out of order. Listwise: It is similar to pairwise. It reduces loss function. Finally, there are two measures which are usually used to assess the effectiveness of retrieval method. The first one is called precision rate which is equal to retrieved relevant document that is actually retrieved. Secondly, recall rate. It is equal to retrieved relevant documents that are actually successfully relevant. If we want to raise precision, then we have to narrow queries. If we want to increase the recall, then we broaden the query. So the relation between precision and recall is an inverse relation. F-measure is the third one that combines precision and recall.

The rest of this paper is organized as follows: modeling of information retrieval systems in section 2and conclusion in section 3.

2. Modeling of information retrieval system's

The main two problems in vector space model are assuming terms are independent and term weighting is intuitive but informal as shown in Figure 3.

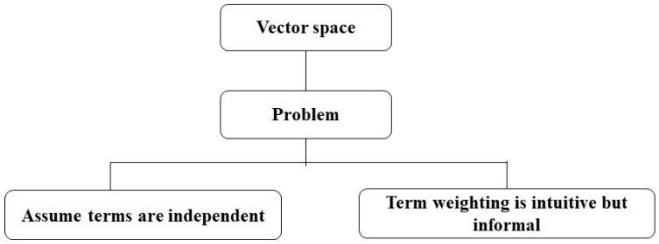

Figure 3. Problems of vector space

2.1 Term dependency modeling

There are many models that assume terms are independent such as Binary independence language (BIL), ontology..etc. Now, dependency modeling is discussed. In Figure 4, the dependency modeling is shown.

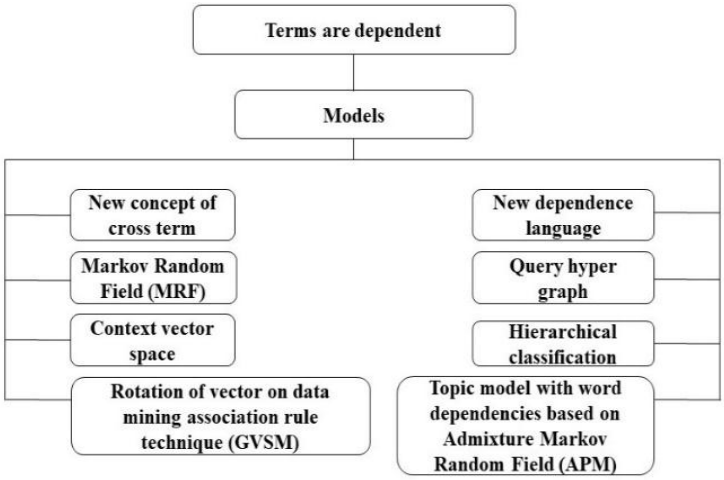

Figure 4. Dependency modeling

In [1], a new concept of the cross term with bigram, n-gram and kernel functions were proposed. Firstly, Cross Term with bigram was used as Basis, then n-gram was generalized for n queries such that (n>=3). Example of bi-gram is shown in figure 5. Shape functions are used to describe the impact of the query term. The functions should be satisfied proprieties (non-negative, continues, symmetric, monotonic and identified). Shape function has seven kernel functions that satisfy effect of query term such as (Gaussian kernel, triangle function, circle function, cosine kernel, Quadratic kernel, Epanechnikov kernel and triweight kernel). The Gaussian kernel was widely used especially in statistics and machine learning. Triangle, circle, and cosine were used in genomic graphics. These functions were used to find the distance. If the distance is raised, then the effect is weak. All of these models were applied on (CRTER2).

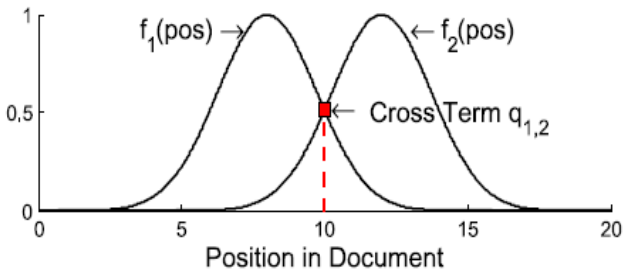

Figure 5. An example of bigram cross term [1].

The shapes of several equations are following:

Gaussian kernel:

$$\text{kernel}\,(u) = \exp\left|\frac{-u^2}{2\sigma^2}\right|, \tag{2}$$

Triangle kernel:

$$\text{kernel}\,(u) = \left(1 - \frac{u}{\sigma}\right).1_{\{u\leq\sigma\},} \tag{3}$$

Circle kernel:

$$\text{kernel}\,(u) = \sqrt{1 - \left(\frac{u}{\sigma}\right)^2}.1_{\{u\leq\sigma\}}, \tag{4}$$

Cosine kernel:

$$\text{kernel}\,(u) = \frac{1}{2}\left[1 + \cos\left(\frac{u\pi}{\sigma}\right)\right].1_{\{u\leq\sigma\}}, \tag{5}$$

Quadratic kernel:

$$\text{kernel}\,(u) = (1 - (\frac{u}{\sigma})^2)^2.1_{\{u\leq\sigma\},} \tag{6}$$

Epanechnikov kernel:

$$\text{kernel}\,(u) = \left(1 - \left(\frac{u}{\sigma}\right)^2\right).1_{\{u\leq\sigma\},} \tag{7}$$

Triweight kernel:

$$\text{kernel}\,(u) = (1 - (\frac{u}{\sigma})^2)^3.1_{\{u\leq\sigma\}}, \tag{8}$$

Dataset of six standard of TREC collection differ in content and size was used. It was measure performance not accuracy and effectiveness.

Bag of words, Bi-term and many dependencies were introduced in, [2]. Some bi-term models exist such as (Bag of Words, Markov random field, Divergence from Randomness and BM5 Model) and many term dependency model such as (BM25-Span Model, Positional language model (PLM)). All of them assume that terms are dependent. BM25-Span model evaluates each span by distorting between width and number of the query term in each span. PLM determines occurrence if each query in the document to neighbor locations. Then kernel functions are used to determine the frequency of each term in Document. The kernel that used is Gaussian kernel functions. Three TREC collections [Robust-04, GOV 2, and Clueweb-09-cat-B] were applied. It discusses performance and effectiveness and determine which is the best in performance and effectiveness but not discuss accuracy, recall, precision not calculated as well.

In [3], a new dependence language extends to language based on unigram was used. Most dependencies do not lead to a development in effectiveness in retrieving large things. There are two reasons for this: - Firstly, the difficulty to estimate dependencies in large scale. Secondly, the integration of both single words and dependencies in weighting schema. The bi-gram language model is better than unigram model. New model expresses term dependencies as a cyclic, planar and an undirected graph. The query is generated from documents in two stages: - Linkage generated is a term generated. This study is applied on different six collection dataset on different models such as binary independence retrieval (BIR), unigram (UG), dependence model (DM), bi-gram language model(BG) and bi-term language model (BT1 and BT2). These comparisons lead to improvement in precision. This is applying to TREC dataset. Comparison with different models "UG", "BIR", "DM", "BI", "BG". Then DM with UG and BIR improve precision effectiveness. But not discuss recall, f-measure, time, accuracy and performance also term weighting.

In [4], Query hyper graphs have also been used to capture complex dependencies between query concepts statements. Queries are acted as vertices and edges. The distance between edges is called "dependencies". Vertex is corresponding to query. Query hyper graph is derived a ranking function that treats with concept and dependency concept. The model is proposed in this research integrates three main characteristics: Model arbitrary term dependencies as a concept, it uses a passage level evidence to model dependencies between these concepts and It assigns a weight to both concepts: concept dependencies and weight to important features. Experiments use newswire and web corpora collection. This framework improves the effectiveness of several retrieval models. Newswire and web corpora collections were used. It discussed effectiveness not performance, accuracy, recall and precision.

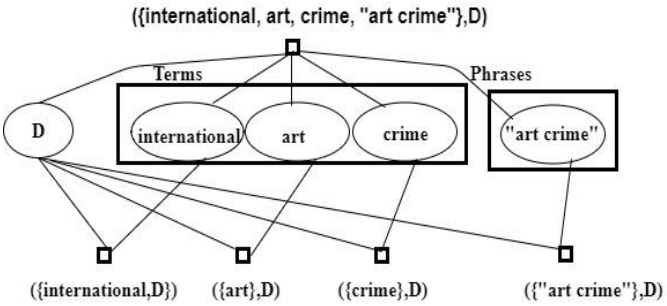

Figure 6. An example of a hyper graph representation for the query "international art crime", [4].

Markov Random Field "MRF" has been used in [5, 6] with a new variation in, [21] to scout full independencies, sequential dependencies, and full dependencies. In [5], MRF is constructed from the undirected graph. Nodes represent random variable and edges define independence between variables. Steps of MRF are to Construct graph for query term dependencies and define a set of potential function and rank document. Full independencies (FI): Terms which occur and don't affect by other terms. Sequential dependencies (SD): terms are dependencies between neighbors. Full dependencies (FD): Terms which depend on each other and show a complete graph and capture longer range dependencies.

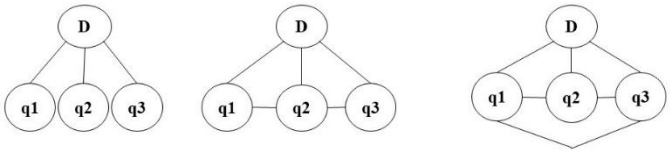

Figure 7. An example of MRF for three query term dependencies (left "FI", middle "SD", right "FD") [5].

The potential function is very important to complete efficiency. The good potential function is assigned a high value to most compatible with each other under given distribution our result shows that modeling dependencies improve effectiveness across a range of TREC collection. Sequential dependencies using ordered features are more effective on the smaller collection while the full Ones are the best for larger.

$$\log \varphi_t(c) = \lambda_t \log p(q_i \backslash D) = \lambda_t \log[(1-\alpha_D)\frac{tf_{qi,D}}{|D|} + \alpha_D \frac{cf_{qi}}{|C|}] \quad (9)$$

Where |D| is the total number of terms in documents, |C| is the length of collection, $tf_{(qi,D)}$ is the number of times occur in document, cf_{qi} is the number of times occur in entire collection and α is the smoothing parameter. TREC, WT10g, and GOV2 were applied in [5]. Full dependencies and sequential dependencies make improvement in average precision rather than full independencies but not discuss the recall and weighting of terms. In paper [6], both publicly available TREC corpora and proprietary web corpus were applied. It discussed effectiveness but didn't discuss performance and accuracy. In [21], a new variation of Markov Random Field that relies on BM-25 was proposed. MRF is one of dependency models. It takes the most attention in recent years. It gives a clear effectiveness not performance because of high computational costs. The new model is applied on TREC8, GOV2, Clueweb09-Category-B collections. It reduces costs by up 60% and keeps the effectiveness as the same with no loss.

In [7], a new topic model based on an admixture of Poisson Markov Random Field (APM) was introduced. APM model dependencies between words are opposed to previous independent topic model such as PLSA. Poisson MRFs (PMRFs) provides JOINT distribution over multivariate count data. The model PMRF (θ, Θ) is defined as follows:

$$pr_{PMRF}(x\backslash\theta,\Theta) \propto \exp\{\theta^T x + x^T \Theta x - \sum_{s=1}^{p} \ln(x_s!)\} \quad (10)$$

Based on the dependency of parameter Θ, if Θ is negative then the dependency is rarely co-occurred. Else, often co-occur. The experiments are applied on Grolier encyclopedias that provide visually appealing and interpretable Results.

In [8], the model is an extension of the vector space model using the association rule of data mining techniques to discover set of terms that co-occur in the document collection. The advantages of vector space model are partial matching, good ranking, simple, fast and it is the most used definition of term weight. In this research, incorporate information of correlation among terms in the collection of vector space model improves effectiveness. Generalized vector space model (GVSM) is another extension of vector space model. In GVSM, terms can be non-orthogonal and represented by a small component called "minterms" with binary weight. In VSM evaluates the degree of similarity between query and documents based on distance. The similarity between two vectors is calculated as follows:

$$\mathrm{sim}(d_j, q) = \frac{d_j \cdot q}{|d_j| \times |q|} = \frac{\sum_{i=1}^{t} w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \quad (11)$$

And the confidence equation is:

$$\text{confidence}\ (A \Longrightarrow B) = p(B/A) = \frac{\text{freq}(A \cap B)}{\text{freq}(A)} \quad (12)$$

A priori for association rules represents a frequent pattern of data. This algorithm identifies co-occurrence among terms. Vector space with association rule is the confidence of terms. Confidence chooses to compute the new angle between term vectors. Term vectors are brought close together according to rules. If 90 then orthogonal angle between vectors K_i, K_j then rotation occurs only in K_i, K_j not modified. The angle is given by:

$$\theta_{ij} = 90(1 - c_{ij}) \quad (13)$$

Where θ_{ij} is a new vector between(K_i, K_j) and c_{ij} is a confidence of association rule$K_i \longrightarrow K_j$. The experiments were made with four reference collections named CACM, CYSTIC FRIBOSIS (CFC), CISI and third text retrieval conference (TREC-3). The result shows the effectiveness of retrieval. The result with association rules is better rather than VSM and improves precision leading up to 31%. This model measures effectiveness in four collections. Not measure accuracy, time and weighting and also not used large dataset.

In [9], a novel method of question retrieval "hierarchical question classification" was used. People can ask the question with natural language instead of segment words and answer to question posted online or offline. Question retrieval is very important which attracts a great deal of interest from the research and community. Overview of question retrieval is illustrated in the Figure 8. Experimental result on Yahoo dataset shows the effectiveness of this method compared to another state of art method. This method is applied by using yahoo answer dataset. It measures performance of retrieval and accuracy but doesn't discuss recall, precision-measure.

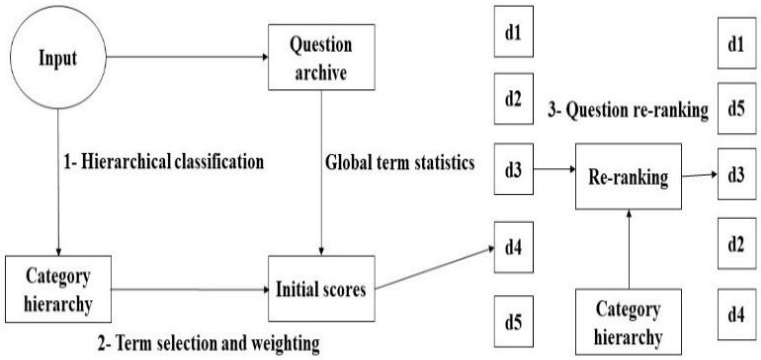

Figure 8. Question retrieval workflow.

Term context vector model was used in [10], based on co-occurrence of term in the same document. Vectors are used to calculate context vector for a document. A context vector is obtained from words occurring close to an entity in a text. Vectors represent the context of single occurrence. Set of term context vector can be represented by nxn matrix as follows:

$$\begin{pmatrix} c_{11} & \dots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

The use of term context vectors allows two different interpretation of index:- syntactic interpretation and semantic description. Once a term correlation matrix has been generated, an initial document into context vector is released. The following equation making is as shown below:

$$\vec{d} = \frac{\sum_{j=1}^{n} w_{ij} \frac{\vec{t_j}}{|\vec{tj}|}}{\sum_{j=1}^{n} w_{ij}} \quad (14)$$

This model performs well in long documents and mysterious query such as medical (MED). It gives high performance and works very well in document with high recall only. Also applying it on CISI, CACM (short documents, specific information for query). It gives better result in MED because different characteristics of information but doesn't measure accuracy, effectiveness, weighting, dependencies.

In [23], extension of relevance model (RM) with the context information was proposed. New of this method is that merge feedback through document language models improvement. Feedback comes from different two ways, firstly, it comes from relevant document, secondly, comes from non-relevant document. Traditionally relevance model (RM) incorporates relevant document which used to estimates and improves query language model. But this method is differing from traditional RM such that it depends on non-relevant documents. Text retrieval conference (TREC) was used to test this method. Retrieval performance could be enhanced by estimating maximum likelihood query model using the following equation:

$$p(w\backslash\theta_Q) = \alpha\, p(w\backslash\hat{\theta}_Q) + (1+\alpha) p(w\backslash\hat{\theta}_Q) \quad (15)$$

Where α: mixing factor.

$$p(w\backslash\hat{\theta}_Q) = \frac{tf(w,Q)}{|Q|} \quad (16)$$

Where $tf(w,Q)$: is the number of occurrences of word w in the query Q, |Q|: total number of words in the query. And the probability equation using maximum likelihood estimate:

$$p(w\backslash\hat{\theta}_Q) \propto \frac{1}{N_{RF}} \sum_{D\in R} \frac{tf(w,D)}{|D|} \quad (17)$$

Where N_{RF}: number of relevant judgement in RF. The context dependent relevance document (CDRM) calculates document ranking score through the following equation:

$$S_{CDRM}(Q,D) \propto \sum_{w\in Q} s(w) \log\left(\max\left(P_{BD}(w\backslash\hat{\theta}_D), \varepsilon\right)\right) + \sum_{w\epsilon Q_{QE\backslash Q}} s(w) \log\left(P(w\backslash\hat{\theta}_D)\right) \quad (18)$$

Where ε: small positive number, $\log(P(w\backslash\hat{\theta}_D)$: maximum likelihood, second sum: is over all words in the expanded query Q_{QE}. CDRM contain various parameters and (x) denotes that unigram (u) or bigram (b) show in the following table 1:

Table 1: Summary of Parameters

symbol	description
N_{RF}	Number of relevant judgments made in RF
N_{QE}	Number of query expansion terms selected from judge relevant documents
α	Mixing factor in the RM3 query expansion that used maximum likelihood query model
${C^{(x)}}_B$	Size of context in known relevant documents for extracting boost terms
${C^{(x)}}_D$	Size of context in known irrelevant documents for extracting discount terms
${\gamma^{(x)}}_D$	Logistic coefficient for discount terms equation
df_B	Document frequency threshold for boost bigram pruning
df_D	Document frequency threshold for discount bigram pruning

Semantic weighted dependence model (SWDM) and pseudo relevance feedback (PRF) was proposed in [24] based on query expansion method. SWDM is reduce mismatch between queries and documents through query expansion. It is look like sequential dependence model (SDM). The score of retrieved documents depends on matching query unigram, ordered and unordered bigram. Unlike SDM and SWDM find the closest unigram and bigram to query terms in embedded space and merge them in to retrieval of SWDM. In word embedding use cosine similarity between them. Graphical representation of SWDM is show in Figure 9. It is applied in two different ways. Firstly, it is used for calculating similarity to find terms that are semantically similar to query terms for query expansion. Secondly, it is used as features to calculate importance of query concepts and it is used extension of SWDM. Dataset is applied on SDM, WSDM, EQE1 and SWDM. The result is improving SWDM over WSDM and EQE1. It calculates mean average precision and it discuss accuracy but not discuss effectiveness and recall.

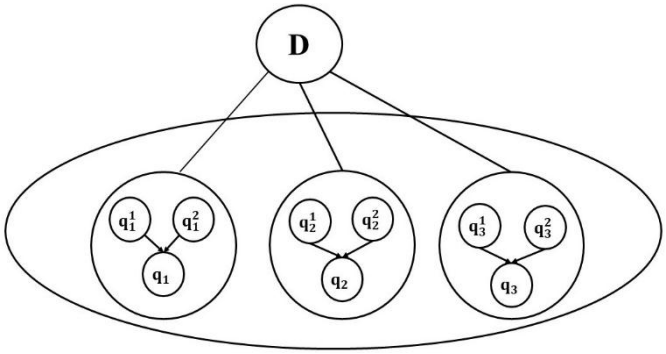

Figure 9. Graphical representation of SWDM.

2.2 Term Weighting Methods

Term weight is another factor that affects the result. There are different methods to measure weighting. These are classified as supervised (TFIDF) and unsupervised (Gain Ratio and Confidence weight) that is shown in Figure 10.

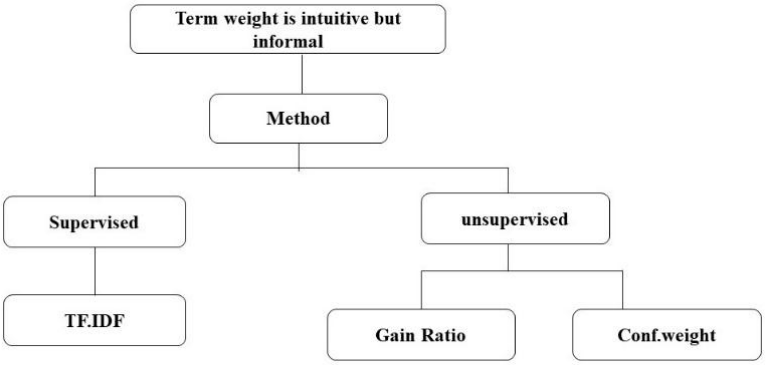

Figure 10. Term weighting methods

It can improve retrieval performance. Many different terms weighting schemes have been proposed in, [10, 11, and 12]. In [12], term frequency (TF) and inverse document frequency (IDF) are the most used and most effective weighting scheme. Term frequency is based on an occurrence of the term in a document. Inverse document frequency is based on a term which occurs in many documents. The TF*IDF is calculated as follows:

$$\mathrm{idf}(t_i) = \log \frac{N}{n_i} \tag{19}$$

$$w_i = tf_i * \log \frac{D}{df_i} \tag{20}$$

Where D is the number of documents, t_i is the term count and df_i is the number of documents containing term (i).It is applying on the standard dataset and found that TF*IDF is high relevance. Deviation of term weighting is discussed in, [10]. Terms that occur in all documents with a different frequency are more important than terms occur in all documents once. A Kind of deviation assesses that different occurrence should improve retrieval performance. The experiments are applied to four different collections (MED, CRANFIELD, CISI, and CACM). The variation of TF*IDF standard across document vector shows as follows:

The modified average mean deviation of term context vector:

$$\mathrm{tcvmamd}(t_i) = 1 + \frac{\sum_{j=1}^{n} \left|\frac{c_{ij}}{\text{mean } \bar{t}_j} - 1\right|}{n} \tag{21}$$

The modified variance of term context vectors:

$$\mathrm{tcvmamd}(t_i) = 1 + \frac{\sum_{j=1}^{n} \left(\frac{c_{ij}}{\text{mean } \bar{t}_j}\right)^2}{n-1} \tag{22}$$

The combination of idf and tcvmamd:

$$\mathrm{tcvmamd}(t_i) = 1 + \mathrm{idf}\,(t_i) \frac{\sum_{j=1}^{n} \left|\frac{c_{ij}}{\text{mean } \bar{t}_j} - 1\right|}{n} \tag{23}$$

In paper [11], Bag of word in order to balancing recall and precision however size of collection continue to increase. Proximity based ranking studied (variant of classical models "OKAPI BM 25, KL- divergence", term dependency model "MRF" "FI, SD, FD"). FD is more suitable on large and less homogenous collection with short queries while SD is more suitable on small and homogenous collection with long queries. Use features based on term applied to SD, FD and BOW. It applied on three of data set TREC (TREC8, Gov 2, and Clueweb09A). It discusses effectiveness and performance but doesn't argue accuracy. In [12], TF*IDF technique for calculating value. It used Standard Test Data. It is High Relevance. Not discuss performance, accuracy, effectiveness.

A variation of standard TF*IDF can be found in, [13]. Term weighting is split into three categories: local, global and normalization. Local weight is calculated according to a number of occurrence terms in document or query. Global weight is a number of occurrence terms in the entire collection. Normalization is done after local and global weight. It is not necessary because it doesn't affect on ranked document list. List of establishing local weight formulas is given in table 2, table 3, and table 4:

Table 2. Local Weight Formula [13]

Formula	Name	Abbr
$1 \quad if \quad f_{ij} \succ 0$ $0 \quad if \quad f_{ij} = 0$	Binary	BNRY
f_{ij}	Within document frequency	FREQ
$1+\log f_{ij} \succ 0 \quad if \ f_{ij} \succ 0$ $0 \quad if \ f_{ij} = 0$	log	LOGA
$\frac{1+\log f_{ij}}{1+\log a_{ij}} \quad if \ f_{ij} \succ 0$ $0 \quad if \ f_{ij} = 0$	Normalized log	LOGA
$0.5+0.5\left(\frac{f_{ij}}{x_{ij}}\right) \quad if \ f_{ij} \succ 0$ $0 \quad if \ f_{ij} = 0$	Augmented normalized term frequency	ATF1

Table 3. Global Weight Formula [13]

Formula	Name	Abbr
$\log\left(\frac{N}{n_i}\right)$	Inverse document frequency	IDFB
$\log\left(\frac{N-n_i}{n_i}\right)$	Probabilistic inverse	IDFP
$1+\sum_{j=1}^{N}\frac{\frac{f_{ij}}{F_i}\log\frac{f_{ij}}{F_i}}{\log N}$	Entropy	ENPY
$\frac{F_i}{n_i}$	Global frequency IDF	IGFF
1	No global weight	NONE

Table 4. Normalization Factors [13]

Formula	Name	Abbr
$\frac{1}{\sqrt{\sum_{i=0}^{m}(G_i L_{ij})^2}}$	Cosine normalization	COSN
$\frac{1}{(1-slope)+slope\, l_j}$	Pivoted unique normalization	PUQN
1	None	NONE

Dataset was used (MED, CISI, CRANFIELD). This apply different formula of term weighting to measure performance but not discuss dependency, accuracy. Also at point of writer different formula achieve efficiency if applied on clustering, phrases and expansion but not applied yet. This is the consideration of future work.

The term weighting is based on question retrieval model shows the relationship between terms pairs when calculating their weight. To overcome this problem, novel term weighting scheme by incorporate dependency relation between two pairs is proposed in, [14]. First, construct dependency graph and compute relation strength between each term. Second, refine initial term weight. The undirected graph ensures that every term pair has a dependency relation path. If the path is shorter, then it reflects stronger relation shown in the Figure 11. This is applied on large dataset of yahoo. It measures performance but not argue dependency, effectiveness and accuracy.

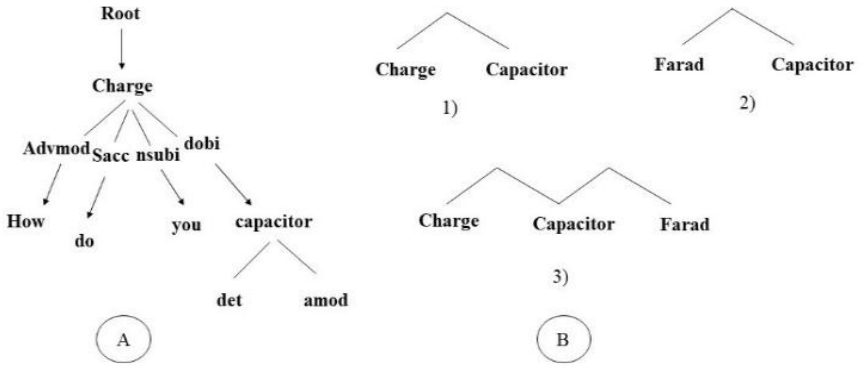

Figure 11. a) Dependency parsing tree and b) Dependency relation path [14].

A new method of term weighting is introduced in, [15]. This method is called confidence weight. It is based on the statistical estimation of the importance of the word. It also has benefit making feature selection. Confidence weight is used as an alternative to TF*IDF. It has the ability to perform well if no features are conducted. If the feature is irrelevant, then weight it gets from confidence weight is low. Three data sets are used to test such as reuters that made of categories related to the business news report, ohsumed that comes from large text collections and rarely used with all available categories and documents of term weight are shown in Table 5.

Table 5. Comparison between Methods

Method	Stable	Sensitive	Accuracy
Gain information ratio	More	Less	Less
Confidence weight	More	More	More
TF*IDF	Less	More	Less

New method "confweight" was proposed and applied on three collection of dataset (Reuters-21578, Ohsumed and Reuters Corpus Vol. 1). There is a comparison of these methods "TF-IDF, Information Gain Ratio" and new method. Problem is that gain ratio fail to show that supervised is higher than unsupervised. And confweight is behave gracefully both with and without feature selection.

In [16], A new term weighting method was proposed. This method doesn't use information of query but uses similarity information between documents. To map similarity of the cluster into weight, we use information gain ratio (IGR). If amount of information of word in cluster increases after the cluster is partitioned into sub-clusters, then word used to determine the structure of sub-clusters. The cluster consists of two sub-clusters. One sub-cluster is the cluster of retrieved documents which will be partitioned into smaller clusters. Another is the rest of database. Structured Similarity describes relation among documents. The clustering retrieval document is based on IGR are shown as follows in Figure 12.

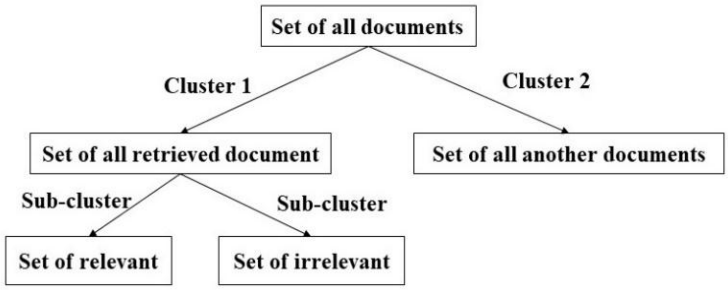

Figure 12. Clustering retrieval documents.

There are three types of term weighting (TF, IDF, and IGR). TF*IDF depends on the distribution of each word in documents. IGR depends on the structure of document cluster and to analyze similarity among retrieved through maximum distance algorithm. The distance is calculated as Euclidean distance as follows:

$$d(D_i, D_j) = \sqrt{\sum_k (weight_{ik} - weight_{jk})^2} \quad (24)$$

$$weight_{ik} = tf(D_i, w_k) idf(w_k) \quad (25)$$

$$tf(D_i, w_k) = \frac{freq(D_i, w_k)}{|D_i|} \quad (26)$$

$$idf(w_k) = \log_2 \frac{N}{df(w_k)} \quad (27)$$

Where N is the total number of documents, $|D_i|$ is the number of morphemes in D_i, freq (D_i, w_k) is the frequency of the word w_k in D_i and df (w_k) is the document frequency of the word.

IGR is the measure used in decision tree learning algorithm C4.5 to choose the best attribute. The dataset used is TSC. It has 12 topics. Each topic has one query and 50 retrieved. After that improving accuracy, time and effectiveness were found. This paper discussed accuracy, time and effectiveness. But the measurement is done separately, there is no certain measure that integrate them in one value. When number of words increase, then it used to determine structure of sub-clusters.it use decision tree to determine best attribute (root of information).

Clustering is unsupervised learning technique. It helps to group data into classes. The cluster is a collection of objects that similar within the same cluster and dissimilar in another cluster. One of the most popular techniques is k-mean algorithm. It is found in [17]. Document clustering is one of the fastest growing research. It is unsupervised learning technique that helps to organize similar document in classes to improve retrieval. Before applying this algorithm, there is preprocessing in [18] should be done such as (tokenization, preprocessing and feature extraction). This algorithm is applied on 20_new group dataset. In [19], Attributed k-mean method task is to map distribution attribute in the applied on 20_new group dataset. In [19], Attributed k- mean method task is to map distribution attribute in the dataset. There are three measures used to evaluate the quality of clustering. First, F-measure: It combines between recall and precision. Secondly, Entropy: it is used to measure the goodness of unnested cluster. It determines how homogenous cluster is. So the relation between homogeneity and entropy is inverse. Higher homogeneity has lower entropy.

$$\mathrm{E}(\mathrm{c_i}) = -\frac{1}{\log \mathrm{c}} \sum_{\mathrm{h}-1}^{\mathrm{k}} \frac{{\mathrm{n_i}}^{\mathrm{h}}}{\mathrm{n_i}} \log\left(\frac{{\mathrm{n_i}}^{\mathrm{h}}}{\mathrm{n_i}}\right) \quad (28)$$

Where c_i is the cluster and n_i is the size. Thirdly, purity: it evaluates the coherence of cluster. To achieve high quality, maximize F-measure, purity and minimize entropy.

$$\mathrm{p}(\mathrm{c_i}) = \frac{1}{\mathrm{n_i}} \max_{\mathrm{h}}({\mathrm{n_i}}^{\mathrm{h}}) \quad (29)$$

Weighted k-mean was proposed. This is applied on mininewsgroup20. It was used to esteem clustering performance on Recall, Precision, f-measure and time. Not discuss dependency, accuracy and effectiveness. Traditional TF-IDF considers the weight of term frequency and inverse document frequency not a concern with the weight of other feature of the word. A new method called TF-IDF adaptive position in [20] determines the weight of position of the word. TF-IDF-AP is applied on Chinese expression. F-measure of TF-IDF-AP has improved by 12.9% compared with classical TF-IDF. It applied on Chinese words. It measured recall, precision and f-measure and find new method is improved with 12.9% rather than classical TF_IDF but didn't discuss dependency, accuracy and performance.

The formula for the position of the first occurrence described as:

$$\text{First position}(\text{word}) = \frac{\text{FPBeforecount (word)}+1}{\sum_{\mathrm{i}=0}^{\mathrm{n}} \text{count } (\text{word}_\mathrm{i}) - \text{FPBeforecount (word)}} \quad (30)$$

The formula for the position of last occurrence described as:

$$\text{Last position}(\text{word}) = \frac{\text{LPAftercount (word)}+1}{\sum_{\mathrm{i}=0}^{\mathrm{n}} \text{count } (\text{word}_\mathrm{i}) - \text{LPAftercount (word)}} \quad (31)$$

The formula for Adaptive weight described as:

$$\text{Position weight} = \frac{1}{\text{First position (word)} + \text{Last position (word)}} \quad (32)$$

The formula of weight described as:

$$\text{weight (word, Doc)} = \frac{\text{TF} * \text{IDF} * \text{position weight}}{\sqrt{\sum_{\text{word,Doc}} [\text{TF} * \text{IDF} * \text{position weight}]^2}} \quad (33)$$

3. Conclusion

At the end of this survey, we conclude that, information retrieval systems (IRS) are used in different areas. It also describes classification of mathematical modeling and term weighting methods for dependency terms. The goal of this survey is to discuss models in detail. This paper explains and compares these models, their advantages, disadvantages and why there is need to use in IR.

References

[1]. JIASHU ZHAO, JIMMY XIANGJI HUANG, and ZHENG YE; "Modeling Term Associations for Probabilistic Information Retrieval"; ACM Trans. Inf. Syst. 32, 2, Article 7 (April 2014), 47 pages.

[2]. Samuel Huston and W. Bruce Croft; "A Comparison of Retrieval Models using Term Dependencies"; ACM; November 3–7, 2014, Shanghai, China.

[3]. JianfengGao, Guangyuan Wu; "Dependence Language Model for Information Retrieval"; special interest group of information retrieval "SIGIR"; ACM; 2004.

[4]. Michael Bendersky, W. Bruce Croft; "Modeling higher order term dependencies in information retrieval using query hyper graph"; special interest group of information retrieval "SIGIR", ACM; USA; 2012.

[5]. Donald Metzler, W. Bruce Croft ; "Modeling Query Term Dependencies in Information Retrieval with Markov Random Fields"; special interest group of information retrieval "SIGIR" ; 2005.

[6]. Michael Bendersky, Donald Metzler, W. Bruce Croft; "Learning Concept Importance Using a Weighted Dependence Model"; ACM; 2010.

[7]. David I. Inouye, PradeepRavikumar, Inderjit S. Dhillon; "Admixture of Poisson MRFs: A Topic Model with Word Dependencies"; International Conference on Machine Learning; JMLR; volume 32; Beijing, China; 2014.

[8]. Silva, I.R. ; Univ. Fed. de Uberlandia, Brazil ; "Dependence Among Terms in Vector Space Model" ; Database Engineering and Applications Symposium; IEEE; 2004.

[9]. Wen Chan, Jintao Du, Weidong Yang, Jinhui Tang, Xiangdong Zhou; "Term Selection and Result Re ranking for Question Retrieval by Exploiting Hierarchical Classification"; ACM; November 03-07 2014, Shanghai, China.

[10]. HolgerBillhardt, Daniel Borrajo, Victor Maojo ; "A Context Vector Model for Information Retrieval"; American Society for Information Science and Technology; vol. 53, n. 3; p. 236-249; 2002.

[11]. Xiaolu Lu, Alistair Moffat, J. Shane Culpepper; "How Effective are Proximity Scores in Term Dependency Models?"; ACM; November 27–28 2014, Melbourne, Victoria, Australia.

[12]. DeepikaMatta, ManojVerma; "Evaluating Relevancy Of Words In Document Queries Using Vector Space Model"; Journal of Engineering, Computers & Applied Sciences (JEC&AS); Volume 2, No.6; ISSN No: 2319- 5606; 2013.

[13]. Erica Chisholm and Tamara G. Kolda; "New term weight formulas for the vector space method in information retrieval"; Computer Science and Mathematics Division; 1999.

[14]. Weinan Zhang, Zhao- yanMing ; "The Use of Dependency Relation Graph to Enhance the Term Weighting in Question Retrieval"; computational linguistics "COLING"; pages 3105–3120; Mumbai; 2012.

[15]. Pascal Soucy , Guy W. Mineau ; "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model "; international joint conference on artificial intelligence "IJCAI"; 2005.

[16]. Tatsunori Mori, Miwa Kikuchi, Kazufumi Yoshida; "Term Weighting Method based on Information Gain Ratio for Summarizing Documents retrieved by IR systems"; Journal of Natural Language Processing;

[17]. Taylor & Francis Group ,"The top ten algorithms in data mining", LLC, 2009.

[18]. Gerald J. Kowalski, Mark T. Maybury,"Information storage and retrieval systems theory and implementation", second edition.

[19]. Monika Gupta, Kanwal Garg;"Attribute Weighted K-means For Document Clustering"; International Research Journal of Engineering and Technology (IRJET), June-2016.

[20]. Jie Chen, Cai Chen and Yi Liang," Optimized TF-IDF Algorithm with the Adaptive Weight of Position of Word"; 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE), 2016.

[21]. Xiaolu Lu, Alistair Moffat, J. Shane Culpepper, "Efficient and Effective Higher Order Proximity Modeling", ICTIR, 16 , September 12–16, 2016, Newark, DE, USA, ACM.

[22]. Goker, A., and Davies, J." Information Retrieval Models", November 2009.

[23]. Edward Kai Fung Dang, Robert W.P. Luk, James Allan." A Context- Dependent Relevance Model" ;journal of the association for information science and technology;2016.

[24]. Saeid Balaneshin-kordan, Alexander Kotov, "Embedding-based.ery Expansion forWeighted Sequential Dependence Retrieval Model"SIGIR , 2017.