

Glossary of IR terms

Elif Aktolga

January 27, 2010

11-Point Interpolated Average Precision A measure for the Precision-Recall Curve reduced to 11 numbers. For each information need, the interpolated precision over all queries is measured at 11 recall levels in $[0.0;1.0]$. Then, for each recall level the arithmetic mean of the interpolated precision is calculated.

20 Newsgroups A standard test collection consisting of 1000 articles from 20 Usenet newsgroups.

Access Control Lists User authorization control by incorporating access control lists (ACL) into the retrieval system. A postings list of documents is retrieved by means of the user name, which is intersected with the actual search result.

Accuracy The fraction of classifications that are correct: $Accuracy = \frac{tp+tn}{tp+fp+tn+fn}$. This effectiveness measure is often used for evaluating machine learning classification problems, however it is not a good classifier for IR because usually over 99% of the documents are not relevant.

Ad Hoc Retrieval The most standard IR task. In order to satisfy an information need, a system is initiated with a user query and documents are provided from the collection.

Auxiliary Index Small index for storing new documents, which is typically kept in memory. A search result is merged with the original index, filtering out outdated results.

Average-link Clustering An agglomerative, hierarchical clustering algorithm that calculates the similarity of two clusters by means of all average similarities between the documents:

$$S(C_1, C_2) = \frac{1}{|C_1| \times |C_2|} \sum_{u \in C_1, v \in C_2} S(u, v)$$

This clustering algorithm has the best effectiveness for IR, avoiding the pitfalls of single-link clustering and complete-link clustering.

Bag Of Words Model A model in which the exact ordering of the words in a document is ignored:

$$c_{map} = \operatorname{argmax}_{c_j \in C} P(c_j) \times \prod_{w \in V_d} P(w_i | c_j)^{N(d, w_i)}$$

where V_d is d 's vocabulary and $N(d, w_i)$ is the number of tokens w_i in d . More concretely, the set of weights for a document d can be viewed as a vector, in which each component represents a distinct term.

Bigram Language Model Also N-Gram Language Model. A language model that conditions on the n previous words:

$$P(w_1 \cdots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2)P(w_4|w_3) \cdots P(w_n|w_{n-1})$$

For the Bigram Language Model, typically we have $n = 2$.

Binary Independence Model Also called ‘BIM.’ A model that has traditionally been used with the *Probability Ranking Principle* (PRP). Here, ‘binary’ means Boolean: documents are represented as Boolean term incidence vectors. A document is a vector $\vec{x} = (x_1, \dots, x_n)$, where $x_i = 1$ iff term i is present in document d . As a consequence, many possible documents have the same vector representation. The model recognizes no association between terms (independence assumption). The ranking function for query terms is as follows:

$$O(R|q, d) = O(R|q) \cdot \prod_{i:d_i=q_i=1} \frac{p_i}{u_i} \cdot \prod_{i:x_1=0;q_i=1} \frac{1-p_i}{1-u_i}$$

where the left product is over query terms found in the document and the right product is over query terms that are not present in the document. Also see ‘Retrieval Status Value.’

Biword Indexes Pairs of consecutive terms in a document as a phrase. Example: The text “Friends, Romans, Countrymen” would generate the biwords ‘friends romans’ and ‘romans countrymen.’ Each of the biwords is treated as a dictionary term.

Blocked Storage Extension of the dictionary-as-a-string data structure, for which terms in the dictionary string are grouped into blocks of size k , keeping a term pointer only for the first term of each block. The length of the term is stored in the string as an additional byte at the beginning of the term. This technique ensures better compression.

Block Merge Algorithm Algorithm to construct an index of large collections. Postings are accumulated in memory until a block of a fixed size is full. This block is then inverted (=sorted) and written to disk. This is repeated for all blocks, which are then merged simultaneously.

Boolean Retrieval Model Queries are in the form of a boolean expression of terms. Terms can be combined with the operators AND, OR, and NOT. Each document is viewed as a set of words.

Break-Even Point A measure for the Precision-Recall Curve, measuring the value at which precision and recall are equal. Like ‘Precision at k ,’ it describes only one point on the precision-recall curve, rather than summarizing effectiveness across the curve. It is usually better to use other measures like the F-measure (for measuring the best point) or the Precision at k or R-measure (for looking at a particular region of the curve).

Case folding Reducing all letters to lower case. This way, terms will match regardless of where and in what form they appear in a sentence.

Centroid A cluster centre. It is defined as the mean or centroid $\vec{\mu}$ of the documents in a cluster w :

$$\vec{\mu}(w) = \frac{1}{|w|} \sum_{\vec{x} \in w} \vec{x}$$

Centroid clustering Also called ‘Representative-based Clustering.’ A clustering algorithm that calculates the similarity of two clusters by means of the similarity of their centroids. The difference between Average-based Clustering and Centroid clustering is that the latter one excludes pairs from the same cluster, whereas the former one considers all pairs of documents in computing the average.

Champion Lists An inexact Top K Document Retrieval Scheme that chooses the set of m documents with the highest tf values for term t . When a query is presented, the cosine similarity is only computed for the union of the champion lists for each of the terms in the query. m should usually be set higher for rarer terms – intuitively $m \geq K$.

Clarity Score Also see ‘Kullback-Leibler Divergence.’ Clarity is the Kullback-Leibler divergence of a model with respect to the corpus. If the distribution of words in a model is identical to that of the entire corpus, it will show very low clarity. Models that are highly focused will have very high scores. Clarity is defined as:

$$clarity = \sum_{w \in V} P(w|Q) \log_2 \frac{P(w|Q)}{P(w|C)}$$

Classification Given, a set of *classes*, seeking to determine which class(es) a given document belongs to. *Standing queries* can be used to classify documents, in which case this process is also referred to as *routing* or *filtering*. The classification of documents into two classes is called *two-class classification*.

CLEF A standard test collection (Cross Language Evaluation Forum), concentrating on European languages and cross-language IR.

Clickthrough Log Analysis Also ‘Clickthrough [Query Log] Mining.’ A measure of the frequency with which people click on the top result in a search result, or any result on the first page.

Cluster A grouping of documents into subsets. Documents within a cluster should be as similar as possible, but documents in one cluster should be as dissimilar as possible from documents in other clusters. Clustering is the most common form of unsupervised learning. Clustering algorithms use the similarity measure for clustering documents, which is usually the *cosine similarity*. Also see ‘Hard Clustering’ and ‘Soft Clustering.’

Cluster Hypothesis states a fundamental assumption: Documents in the same cluster behave similarly with respect to relevance to information needs.

Clustering Algorithms See ‘Single-link clustering’, ‘Complete-link clustering’, ‘Average-link clustering’, ‘Centroid clustering’, and ‘K-means.’

Cluster Pruning An inexact Top K Document Retrieval Scheme that reduces the amount of documents being processed at query time: At preprocessing time, \sqrt{N} documents are randomly chosen from the collection as *leaders*. The nearest *followers* (all the other documents) are computed for these leaders. The cosine similarity is applied here.

Collection Group of documents.

Collection Frequency Total number of occurrences of a term in the corpus.

Combination schemes For common queries on particular phrases (e.g. ‘Michael Jackson’) it is better to use a combination of biword indexes and positional indexes.

Complete-link Clustering A clustering algorithm that calculates the similarity of two clusters by means of the two most dissimilar members from each of them (different cluster-cluster similarity):

$$S(C_1, C_2) = \min_{u \in C_1, v \in C_2} S(u, v)$$

This algorithm produces good clusters, but too few of them. There are a lot of singleton clusters.

Compression See ‘index compression.’

Context-Sensitive Correction A form of spell correction. Unlike ‘Isolated Word Correction’, this method correctly detects misspellings that require understanding of the context (example: ‘flew form Heathrow’). For this, the relative frequencies of biwords in the collection are analyzed.

Corpus A body of texts/documents. Also called ‘collection.’

Cosine Measure For measuring the similarity between two sets A and B: $\frac{|A \cap B|}{\sqrt{|A| \times |B|}}$.

Cosine Similarity Used for measuring the similarity between two documents d_1 and d_2 . The cosine similarity between their vector representations $\vec{V}(d_1)$ and $\vec{V}(d_2)$ is computed as follows:

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)| |\vec{V}(d_2)|} = \frac{\sum_{i=1}^n w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^n w_{i,j}^2} \sqrt{\sum_{i=1}^n w_{i,k}^2}}$$

This is not affected by the document length as the dot product of the two vectors is normalized by the product of their lengths. See also 'Length Normalization.' Hence it holds

$$\text{sim}(d_1, d_2) = \vec{v}(d_1) \cdot \vec{v}(d_2)$$

which is the inner product of the normalized versions of the two documents. The higher $\text{sim}(d_1, d_2)$ is, the more similar are d_1 and d_2 . This way one can search for documents in the collection which are most similar to a specific document. The cosine similarity can also be used to express the score of a document for a query:

$$\text{score}(q, d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{|\vec{V}(q)| |\vec{V}(d)|}$$

Here, a query is regarded as a short document.

Cranfield A standard test collection allowing precise quantitative metrics of IR effectiveness. It contains 1398 abstracts of aerodynamics journal articles, a set of 225 queries, and exhaustive relevance judgments.

δ **Codes** An encoding scheme to store gaps. The *length* part of γ codes is encoded in unary, whereas in δ encoding, it is also encoded as a γ code. Example: 7 is 10,0,11.

δ **Encoding** An encoding scheme to store gaps. The first value in the encoded file is the same as the first value in the original file. Thereafter, each sample in the encoded file is the difference between the current and last sample in the original file. Example: 5 8 10 17 \rightarrow 5 3 2 7.

Dendogram Typical means of visualization for a hierarchical agglomerative clustering. A merge of two clusters is represented as a horizontal line connecting two clusters. The y-axis represents *combination similarity*.

Dictionary-as-a-string Data structure for a dictionary in which the dictionary terms are stored as one long string of characters. For each term its frequency, a pointer to its posting list, and a term pointer are stored. Each term pointer marks the beginning of the new term in the string.

Dictionary Compression Fitting the dictionary into the main memory in order to support high query throughput. Typical data structure: 'Dictionary-as-a-string.'

Distributed Indexing Web search engines use distributed indexing algorithms for index construction, since index construction cannot be performed on a single machine for large collections as the World Wide Web. 'MapReduce' is such an algorithm.

DocID A document identifier, which is assigned to each new document (when encountered for the first time) during index construction.

Document Units that a retrieval system is built over (usually a text).

Document Frequency df_i , the number of documents in the corpus that contain a term t .

Document Likelihood Model A method for using language models in IR. For this, the direction of the ‘Query Likelihood Model’ is simply flipped: From each query q a language model M_q is constructed. The documents are ranked according to $P(q|d)$, where the probability of a query is interpreted as the likelihood that it is generated from the document. Using the Bayes Rule:

$$P(q|d) = \frac{P(d|q)P(q)}{P(d)}$$

Analogous to the ‘Query Likelihood Model’, the results are ranked by $P(d|q)$, which is the probability of the document d being generated from q . This is typically done using the multinomial unigram language (= multinomial Naive Bayes) model:

$$P(d|M_q) = \prod_{w \in D} P(w|M_q)$$

A problem arises with different document lengths, as then the probabilities are not comparable. This model can handle shorter documents better.

Dynamic Summary One of the two types of summaries. Dynamic Summaries are query-dependent, and hence they are customized according to the user’s information need as deduced from a query. They attempt to explain why the document was retrieved for the provided query. Also see ‘Static Summary.’

Edit Distance A form of isolated word spell correction. The aim is to find words which are very similar to the query word. Such words have a low edit distance to the given query word. Given two character strings $S1$ and $S2$, the edit distance (also known as Levenshtein distance) between them is the minimum number of edit operations (insert, delete, replace) required to transform $S1$ into $S2$. Example: The edit distance between cat and dog is 3. Edit operations can also be ‘weighted’ (replacing s by p might be considered more expensive than deleting a).

E-Measure 1– F-Measure (Rijsbergen).

Entropy Measure of uncertainty for a probability distribution:

$$H(p) = H(X) = - \sum_{x \in X} p(x) \log_2 p(x)$$

The more information is available about X , the lower X ’s entropy will be because there will be less surprise about the outcome of a trial. The essential point here is that if a model captures more of the structure of a language, then the entropy of the model should be lower. So entropy can be used as a measure of the quality of the models.

Evaluation in IR For evaluating IR systems and search engines, the following is required:

1. A test document *collection*;
2. A test suite of information needs in the form of *queries*;
3. A set of *relevance judgments*, normally a binary assessment of either *relevant* or *not relevant* for each query-document pair.

An overt expression of an information need is required to be able to judge the test collection set. See also *Precision* and *Recall* for the evaluation of unranked retrieval results and *Precision-Recall Curve* for the evaluation of ranked retrieval results. The aim is to get good recall while having only a small percentage of false positives. Also see ‘11-Point Interpolated Average Precision’, ‘Mean Average Precision’, ‘Precision at k ’, ‘R-Precision’, and ‘Break-Even Point.’

Expectation-Maximization Algorithm Also ‘EM Algorithm.’ An iterative algorithm for model-based clustering that maximizes $L(D|F)$, where F is a set of features. There are two steps, an *expectation step* corresponding to reassignment, and a *maximization step*, corresponding to recomputation of the model parameters.

F-Measure A measure for the Precision-Recall Curve, which trades off precision versus recall. It is the weighted harmonic mean (also see ‘Harmonic Number’) of precision and recall:

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1 - \alpha}{\alpha}$$

The variables are chosen as follows: $\alpha \in [0, 1]$ and $\beta^2 \in [0, \infty]$. The default balanced F measure weights are $\alpha = 1/2$ or $\beta^2 = 1$. It is commonly written as F_1 , which is short for $F_{\beta=1}$:

$$F_{\beta=1} = \frac{2PR}{P + R}$$

Values of $\beta < 1$ emphasize precision, whereas values of $\beta > 1$ emphasize recall. $\beta = 2$ is typically used to weigh recall twice as much as precision, whereas $\beta = 0.5$ is used to weigh precision twice as much as recall. Recall, precision, and the F-Measure are inherently measures between 0 and 1, but they are also written as percentages on a scale between 0 and 100.

Free Text Query A query whose terms are typed freeform into the search interface without any connecting search operators (such as Boolean operators). This style of query is very popular on the web – the query is regarded as a set of terms only.

Front Coding A compression scheme for dictionaries. A common prefix is identified in the first term of a block (*) and then referred to with a special character (#) in subsequent entries. Example: automata, automate, automatic → automat*a, #e, #ic.

γ Codes See ‘ γ encoding.’

γ Encoding A bit-level code to efficiently represent a gap G : a number is encoded as a tuple of *length* and *offset*. *Offset* is the number in the binary representation, but with the leading 1 removed. *Length* refers to the length of the offset in unary code. Example: number: 13 (binary: 1101), offset: 101, length: 1110. → γ code: 1110,101. Advantages of γ encoding: They are prefix-free and parameter-free.

Free-Text Retrieval A query is specified as a set of words without any query operators connecting them.

General Wildcard Queries Given is a wildcard query w as a Boolean query Q on an index, such that the answer to Q is a superset of the set of dictionary terms matching w . Each term in the answer to Q is checked against w and non-matching terms with w are discarded. The result is the dictionary terms matching w , which can now be used with the standard inverted index.

Generative Model A model of language that is used for either recognizing or generating strings (e.g. finite state automata).

Global Document Ordering A global champion list of m documents with the highest values for a term, computed by $g(d) + tf-idf_{t,d}$, where $g(d)$ is a global, query-independent measure for a document (e.g. a number between 0 and 1 based on the number of reviews). At query time, the net scores are only computed for documents in the union of these global champion lists. Thus, documents with large net scores are filtered out and unnecessary computation is avoided.

GOV2 A standard large test collection comprising 25 million web pages.

Grepping Linear scan through documents for keywords (word by word) in linear time.

Hard Clustering computes a *hard assignment* for a document; i.e. each document is a member of exactly one cluster. Also see ‘Soft Clustering.’

Harmonic Number is the sum of the reciprocals of the first n natural numbers (e.g. $\frac{1}{1}, \frac{1}{2}, \frac{1}{3}$ etc.).

Heap’s Law Estimates vocabulary size as a function of collection size: $M = kT^b$ where T is the number of tokens and M is the number of word types in the collection. Typical values for k : $30 \leq k \leq 100$ and $b \approx 0.5$.

Impact A discretized weight by which the importance of a term is represented (other than its frequency).

Index The result of creating an index of term-document pairs in advance in order to avoid linearly scanning texts. It is a binary term-document incidence matrix where the matrix element (i, j) is 1 if the document in column j contains the word in row i , and 0 otherwise.

Index Compression Compression of the dictionary and the inverted index in order to reduce use of space and increase the use of caching. See also ‘Lossy Compression’, ‘Lossless’, ‘Vector Space Model’, and ‘Latent Semantic Indexing.’

Index Elimination An inexact Top K Document Retrieval Scheme that eliminates frequent terms by means of the *idf*. This removes long postings lists of low-*idf* terms.

Indexing Tokenization of text in documents, lemmatization, and construction of an inverted index.

Indirect Relevance Feedback Also called ‘implicit relevant feedback.’ Using indirect sources of evidence to do relevance feedback. This is less reliable, but it is more useful than pseudo-relevance feedback, which contains no evidence of user judgments.

Inexact Top K Document Retrieval Schemes that produce K documents that are likely to be among the K highest-scoring documents for a query. Example schemes: ‘Cluster Pruning’, ‘Index Elimination’, ‘Champion lists’, and ‘Global Document Ordering.’

Information Need The topic about which the user desires to know more, and which is differentiated from a query.

Information Retrieval (IR) finding material (usually documents) of an unstructured nature (usually text) that satisfy an information need from within large collections.

Inverse Document Frequency Measure of informativeness of a term. For this, the weight of a term is scaled so that less relevant terms are filtered. The *idf* of a term t is: $idf_t = \log \frac{N}{df_t}$, where N is the total number of documents in a corpus. The *idf* of a rare term is therefore high, whereas the *idf* of a frequent term is low.

Invert Can be understood as ‘sorting to construct an inverted index.’

Inverted Index A collection of term-postings list pairs. Also called ‘inverted file’.

Inverted List A postings list.

Inverter In the reduce phase of the MapReduce algorithm, the inverter collects all values into a list.

Isolated Word Correction A single query term is corrected at a time even if there is a multiple-term query. Such isolated word correction fails to detect misspellings due to the lack of understanding of the context (example: ‘flew form Heathrow’). There are two methods: ‘edit distance’, and ‘k-gram overlap.’

Jaccard Coefficient For measuring the overlap between two sets A and B: $\frac{|A \cap B|}{|A \cup B|}$. The two sets are k-grams in query q and those in a dictionary term t . The Jaccard Coefficient between the k-gram sets of every dictionary term and the query is computed (1 for same elements, 0 for no match). Only those terms for which the coefficient exceeds a preset threshold (e.g. > 0.8) are added to the output.

Kappa Measure A common measure for agreement between judges (used for relevance assessment). It is designed for categorical judgments and corrects a simple agreement rate for the rate of chance agreement:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of time the judges agreed, and $P(E)$ is the proportion of time they would be expected to agree by chance. The Kappa value is 1 if two judges always agree, 0 if they agree only at the rate given by chance, and negative if they are worse than random. If there are more than two judges, it is normal to calculate an average pairwise kappa value. As a rule of thumb, a kappa value above 0.8 is taken as good agreement, and a kappa value between 0.67 and 0.8 is taken as fair agreement, and agreement below 0.67 is seen as data providing a dubious basis for an evaluation.

K-Gram A technique for handling wildcard queries. A k-gram is a sequence of k characters. Examples: ‘cas’, ‘ast’, and ‘stl’ are 3-grams occurring in the word ‘castle.’ The symbol \$ is used to denote the beginning or end of a word: \$ca, cas, ast, stl, tle, le\$.

K-Gram Index An index with postings lists pointing from a k-gram to all lexicon words containing that k-gram. Example: $etr \rightarrow [metric, retrieval, petrify]$. To search for the wildcard query $re*ve$, we run the Boolean query $\$re$ AND $ve\$$ in the 3-gram index, which yields matching terms such as $relive$, $remove$ and $retrieve$. Then the standard inverted index is employed. A post-filtering step against the original query is required to eliminate non-matching terms that might have slipped in due to a too small k . K-Gram Indexes are also used to determine the edit distance between two strings. Terms with most overlapping k-grams (k-grams in common) are considered similar.

Key-Value Pairs Pairs of term and docIDs (=posting) produced by the MapReduce algorithm.

K-means is the most important flat clustering algorithm. It minimizes the average squared distance of documents from their centroids.

Kullback-Leibler Divergence Also ‘Relative Entropy.’ An asymmetric divergence measure from information theory that tries to measure how different two probability distributions are. In order to compare two different language models, e.g. the query-likelihood model and the document-likelihood model, one would calculate the KL divergence between them as follows:

$$R(d; q) = KL(d||q) = \sum_w P(w|M_q) \log \frac{P(w|M_q)}{P(w|M_d)}$$

Here, a general risk minimization approach for document retrieval is modelled. The KL divergence between d and q is the average number of bits wasted by encoding events from a distribution d with a code based on a not-quite-right distribution q . This quantity is always non-negative, and $D(p||q) = 0$ iff $p = q$. Relative entropy is not a metric – it is not symmetric. Also see ‘Clarity Score.’

Language Identification Written language identification is regarded easier than spoken language identification. A character-level n-gram language identification algorithm was presented by Konheim in 1981.

Language Model A probabilistic language model is a function that puts a probability measure over strings obtained from some vocabulary. Example: A probabilistic finite state automaton, consisting of just a single node with a single probability distribution for each of the different words (also see ‘unigram language model’). The sequence of words does not affect the probability distribution. To find the probability of a word sequence, the probabilities of the individual words simply have to be multiplied:

$$P(\text{frog said that toad likes frog}) = 0.01 \times 0.03 \times 0.04 \times 0.02 \times 0.01 = 0.0000000024$$

Stop probabilities, which are usually required for finite state automata to generate the finite strings, are omitted here.

Among two language models, which give different probability estimates for the words, the model with the higher estimate is more likely to have generated the word sequence. A new language model is estimated by means of a representative sample of text. When a query is presented, the model is used to calculate probabilities of word sequences, after which the documents are ranked and returned. Also see ‘Query Likelihood Model’ and ‘Kullback-Leibler Divergence.’

Laplace Smoothing A method to solve the problem of zero probability by adding 1 to each count:

$$P(w_k|c_j) = \frac{1 + \text{count}(w_k)}{|V| + \text{count}(w_k)}$$

where $|V|$ is the number of distinct terms in the vocabulary.

Latent Semantic Indexing An alternative model to the Vector Space Model that can capture the latent semantic associations of terms. As opposed to the Vector Space Model, this model can deal with synonymy and polysemy (=multiple meanings of a word). For this, singular-value decomposition is used to construct a low-rank approximation C_k to the term-document matrix where k is much smaller than the original rank of C . Each row/column is mapped to a k -dimensional space. As similar terms are mapped to a similar location, similarities between vectors can then be computed.

Leading Wildcard Query A wildcard query with the * symbol at the beginning of the search string. A reverse B-tree is used on the dictionary, in which each root-to-leaf path of the B-tree corresponds to a term in the dictionary written *backwards*.

Lemmatization Reducing the inflectional/variant form of a word to its base form by means of a dictionary and morphological analysis. This results in a proper reduction to the dictionary form of the word.

Lemmatizer A tool that lemmatizes.

Length Normalization To make a unit vector out of a vector by dividing it by its length. The length of a unit vector is 1.

Lexicon A dictionary of terms. Also called ‘vocabulary.’

Linear Zone Combination First generation of scoring methods. Example: $0.6 \cdot \langle \text{Scoring_in_Title} \rangle + 0.3 \cdot \langle \text{Scoring_in_Abstract} \rangle + 0.1 \cdot \langle \text{Scoring_in_Author} \rangle$. Each expression between $\langle \dots \rangle$ takes a value from $[0,1]$. The overall score is in $[0,1]$.

Logarithmic Merging Up to n postings are accumulated in an in-memory index, which is emptied to a new index on the disk when it is full. Only indexes of the same generation on disk are merged (=same number of past merges). Upon search requests both the current in-memory index, as well as the indexes on disk are used. Overall index construction time is $O(T \log T)$ but query processing is slowed by $\log T$ as results from $\log T$ indexes need to be merged.

Lossless Compression A compression technique, in which all information is preserved.

Lossy Compression A compression technique, in which some information is discarded. This makes sense when the lost information is not going to be used anymore. ‘Case folding’, stemming and stop word elimination are forms of lossy compression.

MapReduce Algorithm for indexing on large compute clusters. The map and reduce phases split up the computing job into chunks that are manageable for standard machines. Splits are assigned by the master node to available machines. In case a problem occurs while a machine processes a split, the split is reassigned to another machine. The map function (=parser) produces a list of key-value pairs. All values for a key are then collected into one list in the reduce function. Example: Map: d2: C died. d1: C came \rightarrow [[C,d2],[died,d2], [C,d1],[came,d1]]. Reduce:[[C,d2,d1],[died,d2],[came,d1]]. \rightarrow [[C,d2:1,d1:1], [died,d2:1],[came,d1:1]].

Marginal Relevance Refers to whether a document still has distinctive usefulness after the user has looked at certain other documents.

Master Node Divides up and distributes the work in manageable and reassignable chunks, assigns, and reassigns this for individual worker nodes (=ordinary machines in a cluster). A master node contributes to robust distributed indexing.

Maximum A Posteriori Also called ‘MAP.’ The Naive Bayesian classification principle, in which the posterior probability $P(c_j|d)$ is obtained by multiplying the prior of a class c_j by $P(d|c_j)$:

$$c_{map} = \operatorname{argmax}_{c_j \in C} P(c_j|d) = \operatorname{argmax}_{c_j \in C} P(d|c_j)P(c_j)$$

Maximum Tf Normalization A variant weighting function, in which the individual tf weights for a document are normalized by the maximum tf in that document: $tf_{max}(d) = \max_{\tau} tf_{\tau,d}$, where τ ranges over all terms in d. The normalized term frequency for each term t in document d is defined by:

$$ntf_{t,d} = a + (1 - a) \frac{tf_{t,d}}{tf_{max}(d)^a}$$

where a is a value between 0 and 1 and is generally set to 0.5. It is a smoothing term that damps the contribution of the fraction: tf is scaled down by the largest tf value in the dictionary. Maximum tf normalization is used to reduce higher term frequencies in longer documents, which happens as longer documents tend to repeat the same words.

Mean Average Precision A measure for examining the entire precision-recall curve. For this, precision is averaged when recall increases:

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m} \sum_{k=1}^m \text{Precision}(R_k)$$

If there are several queries with known relevances available, the mean average precision is the mean of the average precisions computed separately for each of the queries.

Mean Reciprocal Rank (MRR) is used in question answering to evaluate the quality of the answers that a QA system returns in response to a question query:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^Q \frac{1}{rank_i}$$

where Q is the set of all questions. The evaluation is based on the mean reciprocal rank of the first correct answer ($rank_i$) returned.

Meta Data Specific data contained in digital documents, with fields such as the date of creation and the format of the document, and often the author and possibly the title of the document.

Multinomial Model A popular model for IR, generating tokens from a document:

$$P(q_1 \dots q_k | M) = \prod P(q_i | M)$$

This model keeps track of the number of occurrences of a token in a document, remembering their positions – just like a counter. Therefore, this model can handle longer documents well. Also see ‘Multiple-Bernoulli Model.’

Multiple-Bernoulli Model A model that generates a binary indicator (0 or 1) for each word w in a document according to its presence (=1) or absence (=0) in the query:

$$P(q_1 \dots q_k | M) = \prod_{w \in q_1 \dots q_k} P(w | M) \prod_{w \notin q_1 \dots q_k} [1 - P(w | M)]$$

Hence, this model ignores the number of occurrences of the words in a document – there is no counter mechanism. As a result, this model works better with shorter documents, as it makes many mistakes for longer ones.

Mutual Information A common feature selection method for two-class classification tasks is to compute the mutual information (MI) of two random variables U and C defined over documents:

$$I(U; C) = P(U, C) \times \log_2 \frac{P(U, C)}{P(U)P(C)}$$

Net Score The net score for a document d is a combination of the global measure of quality $g(d)$, which is query-independent, and a query-dependent score like the cosine similarity:

$$\text{net-score}(q, d) = g(d) + \frac{\vec{V}(q) \cdot \vec{V}(d)}{|\vec{V}(q)| |\vec{V}(d)|}$$

Both components have equal contributions, assuming each is between 0 and 1.

Nibble A 4-bit word.

Normalization Equivalence classing of terms by removing accents, diacritics, using lowercase words only etc. Result: Being able to search for both U.S.A. and USA, for example. Also used in terms of ‘length normalization.’

NTCIR A standard test collection for IR systems. The NTCIR project has built various test collections of similar sizes to the TREC collections, focusing on East Asian language and cross-language IR.

Overlap Score Measure The score of a document d is the sum of all the occurrences of each query term in d . Alternatively, the tf-idf weight of each term in d can be used: $Score(q, d) = \sum_{t \in q} tf - idf_{t,d}$.

Parametric Index An inverted index for some kind of meta-data. For each field, an inverted index is built, whose dictionary consists of all distinct values occurring in that field, and postings point to documents with that field value. Example: For the author field, the dictionary consists of all authors of documents in the collection; the postings list for a particular author consist of all documents with that author. For query processing, a search tree with the ordered values is used.

Permuterm Index A form of inverted index consisting of all rotations of each term (with the \$ terminating symbol appended). Example: [hello\$, ello\$h, llo\$he, lo\$hel, o\$hell] → hello. For example, one wants to search for the wildcard query m*n. The string has to be rotated so that the * symbol appears at the end: n\$m*. This string is looked up in the permuterm index, where it points to the terms ‘man’ and ‘men’. Then the standard inverted index can be employed.

Permuterm Dictionary The set of rotated terms in a permuterm index.

Phrase Queries Queries in double quotes syntax. Example: ‘Stanford University.’ A biword index that includes variable length word sequences is also called a ‘Phrase Query.’

Pivoted Document Length Normalization A form of document length normalization that is independent of term and document frequencies so that the document length is also considered for determining its score. The resulting ‘normalized’ documents are not necessarily of unit length. The inner product score between a (unit) query vector and such a normalized document is computed, and the score is skewed to account for the effect of document length on relevance.

Posting An item in a postings list referring to a docID. The posting indicates which document a term occurs in.

Positional Indexes Posting list entries of the form docID: position1, position2, ... So in addition to looking up documents in posting lists, one also has to check that the positions of appearance of the queries in the document are compatible with the phrase query being evaluated. The positional index size depends on the average document size.

Posting List A collection of docIDs for a term.

Posting List Intersection Merging lists with a logical AND operation. Also called ‘merging two posting lists’.

Postings All postings lists together.

Postings File Compression The docID of the first matching document is recorded only. Compression is achieved by recording the gaps between postings after the first entry since those gaps are short. For an economical representation of the distribution of gaps, a variable encoding method is required that uses fewer bits for short gaps. Example methods: ‘byte-wise compression’ and ‘bit-wise compression.’ Bitwise compression is better if disk space is limited (see also γ codes and δ codes).

Precision Describes the precision of a search result. It refers to the *relevant* collection of documents that were found among all the retrieved documents:

$$Precision = \frac{\#(relevant\ items\ retrieved)}{\#(all\ retrieved\ items)} = P(relevant|retrieved)$$

According to the following contingency table,

	Relevant	Not relevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

precision can also be expressed as follows: $P = \frac{tp}{tp+fp}$. Example: Web surfers require high precision as they want every result on the first page to be relevant. As more documents are retrieved, precision usually decreases whereas recall increases.

Precision At k A measure for the Precision-Recall Curve, measuring precision at fixed low levels of retrieved results for knowing how many good results there are on the first few pages. Example: ‘Precision at 10.’ This has the disadvantage that it does not average well, since the total number of relevant documents for a query has a strong influence on precision at k . Also see ‘R-Precision.’

Precision-Recall Curve A curve for the evaluation of ranked retrieval results. Appropriate sets of retrieved documents are naturally given by the top k retrieved documents (also see ‘Inexact Top K Document Retrieval’). For each such set, precision and recall values can be plotted to a precision-recall curve. The curve is saw-tooth shaped because precision drops rapidly when a nonrelevant document is retrieved. The curve can be ‘corrected’ with interpolated precision at a certain recall level r , which is defined as the highest precision found for any recall level $q \geq r$:

$$P_{interp}(r) = \max_{r' \geq r} p(r')$$

By this interpolation the curve is made to be monotonically decreasing. Also see ‘11-Point Interpolated Average Precision.’

Probabilistic Information Retrieval An information retrieval model which takes a probabilistic approach to ranking documents: they are ranked by the estimated probability of their relevance with respect to the information need. That is, documents d are ordered by $P(R|d, q)$. Also see the ‘Probability Ranking Principle’ and the ‘Binary Independence Model’, which describes how to calculate the required estimates.

Probability Ranking Principle Also ‘PRP’ (by van Rijsbergen). It states that the documents are ranked according to probabilities that are estimated as accurately as possible on the basis of the given data. More concretely, given a document D , its probability of belonging to the relevant class is calculated and it is retrieved if this is greater than its probability of belonging to the non-relevant class (i.e. $P(R|D) > P(NR|D)$). In terms of retrieval costs: Let C_r be the cost of retrieval of a relevant document and C_{nr} the cost of retrieval of a non-relevant document. Then the PRP says that if

$$C_r \cdot P(R|d) + C_{nr} \cdot P(NR|d) \leq C_r \cdot P(R|d') + C_{nr} \cdot P(NR|d')$$

holds, then d is the next document to be retrieved.

Proximity Operator A way of specifying that two terms in a query must occur in a document close to each other, where closeness may be measured by limiting the allowed number of intervening words or by reference to a structural unit such as a sentence or paragraph.

Pseudo-Relevance Feedback Also ‘Blind Relevance Feedback.’ A relevance feedback method that 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, on which relevance feedback can be done. Also see ‘Indirect Relevance Feedback.’

Query Terms that the user conveys to the computer in order to communicate an information need.

Query Expansion An area of IR in which a query is expanded by synonyms etc. e.g. by the use of WordNet or other thesauri.

Query Likelihood Model The original and basic method for using language models in IR. From each document d a language model M_d is constructed. The documents are ranked according to $P(d|q)$, where the probability of a document is interpreted as the likelihood that it is relevant to the query. Using the Bayes Rule:

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)}$$

$P(q)$ is the same for all documents and the prior $P(d)$ is often treated uniform across all d , so they can be ignored. Therefore, results are simply ranked by $P(q|d)$, which is the probability of the query q being a random sample derived from d . This is typically done using the multinomial unigram language (= multinomial Naive Bayes) model:

$$P(q|M_d) = \prod_{w \in V} P(w|M_d)^{c(w)}$$

i.e. documents are ranked by the probability that a query would be observed as a random sample from the respective document model.

Query Optimization The process of selecting how to organize the work of answering a query so that the least total amount of work needs to be done by the system. The standard heuristic here is to process terms in order of increasing term frequency (start with the smallest posting lists). For more complex queries, it is better to intersect each retrieved postings list with the current intermediate result in memory, where the intermediate result is initialised by loading the postings list of the least frequent term.

Ranked Retrieval These systems use a precise language with operators for building up query expressions as opposed to boolean querying which uses free-text queries. Example: Vector space model.

Recall Describes the completeness of a search result. It refers to the amount of the relevant documents retrieved out of all relevant documents available:

$$Recall = \frac{\#(relevant\ items\ retrieved)}{\#(all\ relevant\ items)} = P(retrieved|relevant)$$

Recall measures how well a search system finds what you want, and precision measures how well it weeds out unnecessary items. High recall (=1) can always be achieved by retrieving more (all) documents, however this will lower precision. According to the following contingency table,

	Relevant	Not relevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

recall can also be expressed as follows: $R = \frac{tp}{tp+fn}$. Example: Intelligent analysts require high recall and look at every document that is relevant.

Relevance With respect to a user information need (**not** a query), a document is judged as either *relevant* or *not relevant*. A document is relevant if it addresses the stated information need, rather than containing all the words in the query. The test document collection and the queries to test with must be of a reasonable size so that performance can be averaged over large test sets.

Relevance Feedback Obtaining feedback from the user during the retrieval process in order to improve the final results. This is done by letting the user mark retrieved documents as relevant or not relevant, after which the system returns a better, revised set of results. Also see ‘Pseudo-Relevance Feedback’ and the ‘Rocchio Algorithm.’

Reuters-RCV1 A model collection with roughly one gigabyte of text, consisting of about 800,000 documents.

ROC Curve ‘Receiver Operating Characteristics’. Plots the true positive rate or sensitivity against the false positive rate ($\frac{fp}{fp+tn}$ = how many documents were retrieved that were not relevant), also referred to as $1 - specificity$. A ROC curve always goes from the bottom left to the top right of the graph. For a good system, the graph climbs steeply on the left side (= more correctly retrieved documents, and very few non-relevant retrieved documents).

Rocchio Algorithm is a classic algorithm for implementing relevance feedback. The aim is to separate relevant and non-relevant documents:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

where q_0 is the original query vector, D_r is the set of relevant documents, D_{nr} is the set of non-relevant documents, and α, β, γ are weights attached to each term. That is, the revised query is the weighted vector difference between the centroids of the relevant and non-relevant documents, where we can vary the individual components with the weights: if, for example, there were many judged documents, β and γ would be chosen higher. Reasonable values might be $\alpha = 1, \beta = 0.75$ and $\gamma = 0.15$. Negative term weights are ignored and hence set to 0. Relevance feedback can both improve recall and precision, but it is most useful for increasing recall.

R-Precision A measure for the Precision-Recall Curve, alleviating the problem of averaging badly in ‘Precision at k .’ It requires having an (incomplete) set of known relevant documents of size Rel , from which the precision of the top Rel documents is calculated. This measure adjusts for the size of the set of relevant documents: A perfect system could score 1 on this metric for each query, even if there were only 8 documents in the collection relevant to an information need.

Rule Of 30 States that the 30 most common words account for 30% of the tokens in written text.

Score Of A Document A plausible definition is $\sum_{t \in q} tf - idf_{t,d} - c_1 n_!$, where $n_!$ is the number of exclamation marks in a document and c_1 is a well-chosen constant. Documents with too many exclamation marks are considered to be of low quality.

Segment File A local intermediate file to store the output of a parser.

Semistructured Retrieval XML is a form of semistructured retrieval. Databases for example contain fully structured data, whereas XML documents semi-structure text by means of labeled internal nodes. There is no natural document unit or *indexing unit* in semistructured documents. Thus, given a query it is sometimes unclear which data should exactly be returned. There are three different approaches to defining the indexing unit: one is to index all components that are eligible to be returned in a search result. With this approach the results will inevitably contain overlapping units that have to be filtered out in a post-processing step. Another approach is to group nodes into non-overlapping pseudo-documents. The third approach is to designate one XML element as the substitute for the document unit. Similarly, for computing term statistics (such as idf etc.), different contexts of a term have to be distinguished between, for which typically term-context pairs are used (e.g. a search for the term ‘gates’ is different under the context ‘author’ than it is under the context ‘section.’)

Sensitivity Synonym for recall.

Simple Conjunctive Query Example: Brutus AND Calpurnia.

Single-link Clustering An agglomerative, hierarchical clustering algorithm that calculates the similarity of two clusters by means of the two most similar members from each of them (highest cluster-cluster similarity):

$$S(C_1, C_2) = \max_{u \in C_1, v \in C_2} S(u, v)$$

Skip List Postings lists with skip pointers. Skip pointers are shortcuts that allow to avoid processing unnecessary parts of a postings list. For a list of length P , \sqrt{P} evenly-spaced skip pointers are assigned at indexing time. Building effective skip pointers is easy if an index is relatively static. A skip pointer is used while merging when the end point of one list is still less than the item on the other list.

Smoothing A method to overcome the problem of sparse data caused by the maximum likelihood estimate being based on a particular (small) set of training data. Due to this, the *zero probability problem* arises. Common smoothing methods are ‘Laplace Smoothing’, ‘Dirichlet Smoothing’, and various discounting (lowering) methods.

Snippet A short summary of the document, often provided in search results.

Soft Clustering computes a *soft assignment* for a document; i.e. a document’s assignment is a distribution over all clusters. Also see ‘Hard Clustering.’

Sorting A core indexing step, in which a list of normalized tokens for each document is sorted alphabetically. Such a list consists of pairs of term and (one) docID. An extra column is also added to record the frequency of the term in the document (unnecessary for basic Boolean search).

Soundex An algorithm for phonetic hashing for similar-sounding words. Query and dictionary terms are indexed in a 4-character reduced form together with their original form. The soundex index is then searched when a soundex match occurs. Example: ‘Hermann’ maps to ‘H655.’

Specificity Measures how many of the nonrelevant documents have not been retrieved ($\frac{tn}{fp+tn}$). This measure is not so useful because the set of true negatives is always so large that its value would be almost 1 for all information needs (and, correspondingly, the value of the false positive rate would be almost 0).

Spelling Correction There are two approaches: ‘edit distance’ and k-gram overlap.’

Splits Unit of data (usually web pages) which is split into blocks by the MapReduce algorithm. The size of the split is determined by the computing environment.

Standing Query A query that is periodically executed on a collection to which new documents are incrementally added over time. The query has to be refined over time to achieve good recall. Therefore these queries can gradually become quite complex (like Boolean queries).

Static Summary One of the two types of summaries. It is always the same regardless of the query. Mostly, a set of ‘key’ sentences of the document are used for this. Also see ‘Dynamic Summary.’

Stemming Reducing a word to its ‘stem’ or root by affix chopping (the suffix). The most common algorithm for stemming English is the Porter Stemmer, consisting of 5 phases of word reductions, applied sequentially.

Stop List List of stop words. The general strategy to determine a stop list is to sort the terms by frequency, then including the most frequent terms in the stop list. The terms often must be hand-filtered for their semantic content relative to the domain of the documents being indexed.

Stop Words Very common and semantically non-selective words, which are entirely excluded from the dictionary.

Structured Document Retrieval Principle ‘A system should always retrieve the most specific part of a document answering the query.’ This principle motivates a retrieval strategy that returns the smallest unit containing the information sought.

Sublinear Tf Scaling A variant weighting function. In order to ‘normalize’ tf weights the logarithmic function is used:

$$wf_{t,d} = \begin{cases} 1 + \log tf_{t,d} & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Tf-idf weighting is also changed accordingly: $wf - idf_{t,d} = wf_{t,d} \times idf_t$. Wf can in fact be replaced by any weighting function (example: maximum tf normalization).

Term A ‘word.’ Also ‘token’.

Term-Document Matrix An $M \times N$ matrix whose rows represent the M terms of the N columns (documents). This way, a collection of N documents is viewed as a collection of vectors.

Term Frequency Assigning a weight to each term t in a document d which is determined by the number of occurrences of each t in d . The weighting scheme is denoted $tf_{t,d}$.

Test Collections Standard test collections for ad hoc IR systems evaluation are *Cranfield*, *TREC*, *GOV2*, *NTCIR*, *Reuters RCVI*, *CLEF*, and *20 Newsgroups*.

Tf-idf Weighting A composite weight for each term in a document. The tf-idf weighting scheme assigns to term t a weight in document d given by $tf-idf_{t,d} = tf_{t,d} \times idf_t$. By multiplying the term frequency by the inverse document frequency, the tf-idf measure increases with the number of occurrences of a term within a document and with the rarity of the term across the whole corpus. Hence, terms that are contained in many documents, but fewer times within a document, score low.

Trailing Wildcard Query A wildcard query with the * symbol at the end of the search string.

TREC A standard test collection used by the NIST for IR test bed evaluations since 1992. There have been many tracks over a range of different test collections. The 5 CD TREC Test Collection comprises 1.6 million documents and 450 information needs. There are no exhaustive relevance judgments, as the test document collection is much larger. Rather, relevance judgments are only available for the top k results.

True Casing Case-folding by a machine learning sequence model in order to leave mid-sentence capitalised words as capitalized and lower case all the others.

Type The class of all tokens containing the same character sequence.

Unary Code Simplest bit-level code. The unary code of n is a string of n 1s followed by a 0. Example: number: 9; unary code: 111111110.

Unigram Language Model The simplest form of a *Language Model* that discards conditioning in context (the dependencies of one probability on another) and estimates each word independently:

$$P_{uni}(w_1 \cdots w_n) = P(w_1)P(w_2) \cdots P(w_n)$$

This is a “bag of words” model as the order of words is irrelevant. This unigram language model is the most commonly used one in IR. Other models, such as the *Bigram Language Model*, are rather used in other tasks like speech recognition, where the structure of the sentences plays a role.

Universal Code A code (like γ -code) with the property of being within a factor of optimal for an arbitrary distribution P is called universal.

Variable Byte Encoding A bitwise encoding technique for postings file compression. It uses an integral number of bytes to encode a gap. The last 7 bits encode the size of the gap. The first bit of the byte is a continuation bit. It is set to 1 for the last byte of the encoded gap and to 0 otherwise. To decode a variable byte code, a sequence of bytes is read with continuation bit 0 terminated by a byte with continuation bit 1. Then the 7-bit parts are extracted and concatenated.

Vector Space Model A set of documents as vectors in a vector space, where one axis represents a term. Weights represent various topics that are discussed in a document. Hence, two documents with similar vector representations discuss the same topics. The ordering of the terms is unimportant. Therefore, this model is apt for no-syntax, bag-of-words queries (*free-text retrieval*). It is essential for IR tasks such as scoring documents on a query, document classification and clustering. Similarity between two documents is measured by means of the ‘Cosine Similarity.’ The vector space model can also be applied to queries alone. But tasks like scoring phrases and dealing with wild cards do not suit this approach as a number of *tfs* and *idfs* would have to be computed. Also see ‘Latent Semantic Indexing’ for an alternative model.

Weighted Retrieval Systems, in which postings are often ordered according to weight or impact, with the highest-weighted postings occurring first.

Weighted Zone Scoring Given a Boolean query q and a document d , weighted zone scoring assigns a score in the interval $[0,1]$ to the pair (q, d) by computing a linear combination of zone scores (as Boolean values). Example: The Boolean score from the title zone could be defined as 1 if all the query term(s) occur in the title, and zero otherwise. Or, the Boolean score could be defined as 1 if any of the query terms occurs in the title. Then, the weighted zone score is defined to be the sum of all the scores of the matching items.

Wildcard Query A query that contains at least one * symbol in the search string. Example: mon*. The search tree data structure B-tree is typically used for wildcard queries. The inverted index is then employed with the collected set of W terms in order to retrieve relevant documents. A combination with a reverse B-tree is useful for search strings with the symbol * occurring within the string.

Word Segmentation Used as prior linguistic processing where text is written without any spaces between words.

Zipf’s Law In a corpus, the frequency of any word is roughly inversely proportional to its rank in the frequency table. So, the second most frequent word will have half as many occurrences, the third most frequent word has a third as many occurrences etc.

Zone Identified region within a document. The contents are ‘free text.’ Example: Document titles and abstracts. A separate inverted index is built for each zone of a document, to support queries such as “find documents with *merchant* in the title and *william* in the author list and the phrase *gentle rain* in the body.” \rightarrow william.title, william.abstract, william.author are typical inverted index entries. The size of the dictionary can be reduced by encoding the zone in the postings. Example: william \rightarrow 2. author, 2. title; 3. author etc. This is particularly useful for *weighted zone scoring*.

Note

This glossary was mainly created by extracting definitions verbatim from [1, 2]. A great thanks to the authors for writing such valuable books!

For improvements on this glossary (very welcome!) please email me at [my_first_name]@cs.umass.edu.

References

- [1] Manning, C. D., Raghavan, P., Schütze, H. (2007). *An Introduction to Information Retrieval*. Draft, 1 July 2007, Cambridge UP.
- [2] Manning, C. D., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press.
- [3] Allan, J., Raghavan, H. (2003). *Entity models: Construction and applications*. University of Massachusetts, Amherst, CIIR Technical Report.