Language Models for Intelligent Search Using Multinomial PCA

Wray Buntine, Petri Myllymaki and Sami Perttu

Complex Systems Computation Group (CoSCo), Helsinki Institute for Information Technology (HIIT)
University of Helsinki & Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland.
{Firstname}.{Lastname}@hiit.fi

Abstract. Once one gets beyond the simple keyword searches of internet search engines, Information Retrieval offers more extensive interfaces for searching for relevant documents. Language models for search have been developed recently. In this paper, we show how to modify these models to arrive at something which automatically incorporates learnt knowledge about topics and words in a collection. This requires, however, richer document models than are usually available. Multinomial PCA is the multinomial analogue to Gaussian PCA and is also a natural extension to clustering models but particularly suited to the bag of words model for text. Multinomial PCA turns out to be suitable for the probabilistic task of language modelling. In this paper we present the theory and demonstrate its use on a corpus of 800,000 news articles. While we do not claim this is an ideal approach (for instance, it does not use linguistic knowledge), and our demonstration is not on true HTML data (which is complicated by its poor spelling and large use of named entities), the results are very good it shows that more meaningful search is now realizable. The browser being developed for this research is available now under Open Source.

1 Introduction

Over the last decade there has been enormous growth in statistical natural language processing (e.g., [13]), and since the advent of the Web an explosion of interest in text/HTML/XML based information retrieval and information extraction. *Information retrieval* is generally viewed as the task of "find documents or parts of documents that interest me," where the user supplies a query of some kind to express their interest area. In contrast, *Information extraction* is the more extensive mining for specific things such as addresses, bibliographies or corporate records. *Question answering*, a third area is different again and is to answer a specific question using document contents.

While probabilistic methods caste in a traditional linguistic framework is arguably the dominant paradigm for natural language processing (NLP), and for information extraction, the same cannot yet be said for information retrieval. The

emerging dominant paradigm for question answering is to couple an information retrieval system using NLP essentially for data cleaning with a special purpose theorem prover using a thesaurus as a poor-man's general purpose ontology [15]. A number of frameworks have been proposed for information retrieval, and the most recent probabilistic models (e.g., [11]) are models that include some kind of smoothing or mixing that lies outside the normal rules of probability theory.

Our methodology is to first develop a probabilistic scheme for information retrieval, and then gradually augment it with NLP methods. We report on the first steps here. The probabilistic method provides the mathematical framework for devising a matching algorithm between documents. It should be viewed as complementary to techniques such as NLP or ontology based approaches. The probabilistic method provides the data fusion calculus so that other techniques can be combined with general matching principles.

In this paper we present a model for retrieval based an a probability inference task using a Multinomial PCA model for the document collection [3]. The basic idea is simple: what is the probability that the text of the query would match the text of the document or could match reasonable variations of the same document? That is, we suppose that each document in the collection has some hidden process responsible for generating it and see how well the new query (a few words, a paragraph, or new document) could have also been generated by this same process. Now we do not claim to have a sophisticated model for the generating documents, the "process". The bag of words model is trivial at best, and for instance destroys locality information used in most search engines, and destroys the linguistic clues necessary to disambiguate words. But we do claim that our demonstration of richer query modelling is well on the way to intelligent search (something that cannot be said for keyword search), something badly needed for enterprise information systems and intranets.

2 Retrieval Methods

Retrieval operates through keyword, boolean query or weighted query searches, using an inverted index, or full-text matching which generally requires a full pass of the document database. Hybrids can be done, and are used by some meta-search engines: keyword search produces candidates and full-text matching scores them.

The simple approach of keyword search coupled with all sorts of ranking to boost title words and neighbouring words is remarkably robust. For instance, when coupled with the strategy of only showing the top few matches from a single site, the problem of boiler-plates or templates (for instance, where a menu panel appears duplicated on all pages) is effectively handled. Extensions to the basic keyword approach include query expansion and document expansion. Query expansion means augmenting a query with additional words, or modifying the format. For instance, one might expand "cell phone" to "(cell OR mobile) AND phone". Document expansion does the expansion in the index, and thus has a cost in terms of space. The astute reader might notice that these are not models

in the traditional sense, but implementation schemes. Carefully crafted studies show the approach works well, but the difficult questions are often unanswered. Which expansions should be done, when, and how? A good model should provide the strategy.

Full text matching, on the other hand, requires some kind of distance measure between documents. The tried and true measure in use here is to convert the document to a set of word scores–TF/IDF score or a square-root of frequency–and apply a cosine metric [13]. More generally, measures for text is a field that is still undergoing a discovery phase, with extensive empirical work trying out different methods [5]. These methods sometimes have intuitive probability justifications (e.g., [6]). Considerably more sophisticated probability scores have been developed [7, 9], yet the general task of retrieval has not been addressed as a full inference task. While we believe this is the most promising of current approaches, we take a different alternative here. Our approach is to give a coherent definition of matching documents based on a probability model, implicitly defining a distance metric.

Another approach is to use dimensionality reduction and perform the query by matching in the lower dimensional space, e.g., see the survey in [10]. A popular method is LSI, a variant of PCA. Instead, we can transfer the approach to multinomial PCA, a statistical model shown to provide better dimensionality reduction [8], as is done here. However, the approach fails in a wide range of cases: suppose the query is Hillary Clinton, and only one document in the collection contain these words. Then there will be no trace of "Hillary Clinton" in the statistical model because it appears in the ignored "noise" components. In this case, something akin to keyword search is essential, or the distance metrics considered above which can introduce the effects of rare words.

A final approach is the query model coupled with a *document language model* [16]. Recent approaches offer additional sophistication [11], but the basic idea is to place a generating distribution on queries that is dependent on a document, but also on characteristics of the corpus as a whole. Search then is to return the document with the maximum likelihood. That is, we develop a distribution

$$p(\text{query} \mid \text{document}, \text{corpus})$$

and then we return the *document* maximizing this likelihood. In principle, this allows us to incorporate all sorts of linguistic and corpus information into the score, however proponents have not done so yet. This distribution is usually built by an *ad hoc* mixture the observed frequency for words in the document, and the observed frequency for words in the entire corpus. Given that the observed word frequencies for a document is a poor model at best (as illustrated by the poor performance of standard probabilistic clustering for text [10]), this approach certainly needs improvement. However, it provides a nice starting point.

3 Practical Considerations

We need to point out that the theory here is a small part of achieving a workable solution, though we believe it is important to achieve intelligent search. Repre-

senting documents as bags of words is crude at best. Due to the presence of things like noun compounds, named entities, etc., it becomes very misleading to place all words in bags without some sort of language processing beforehand. Noun compounds induce string dependencies in data that corrupt statistical models of the bag. Moreover, real internet data is intrinsically noisy: gross spelling errors, punctuation and grammatical errors, acronyms, abbreviations and colloquialisms are not uncommon once one moves beyond the realm of mass-circulation professional publications and into intranet data, especially with newsletters and email content.

In the theory below, we consider the "bag of words" model [1], where a document is represented as a sparse vector of word indices and their occurrence counts. All positional information is lost. In practice, we use a "bag of lexemes," after basic processing to identify noun compounds, reduce words to their lemmata, and in the case of commercial websites, remove easily identified template content. All this uses standard techniques available, and makes a considerable difference to results. More generally, additional linguistic processing should be used to perform tasks such as spelling correction, disambiguation, and anaphora resolution, to clean up the quality of the text, and the "bag of lexemes" needs to be replaced so that analysis and retrieval can deal with local context. Our ideal is represented in the diagram in Fig 1. The role of the current paper is

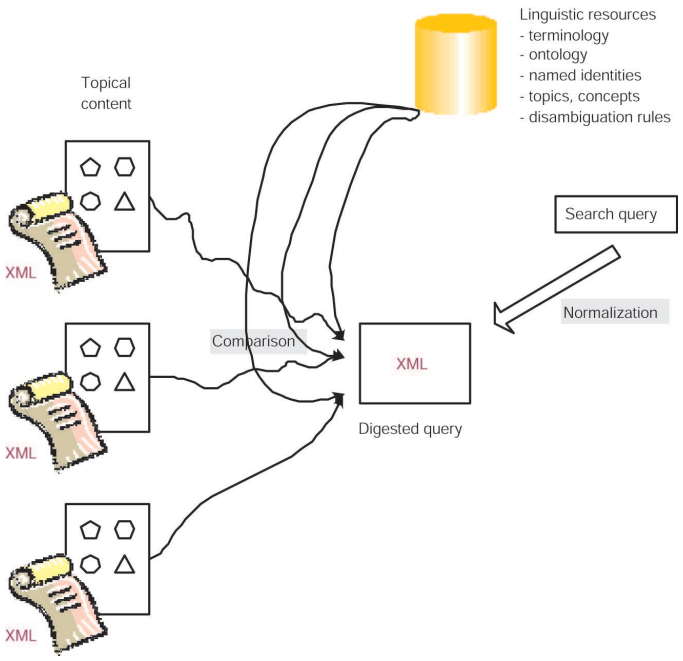

Fig. 1. An ideal XML free-text search function

to demonstrate the improved results one obtains when some small amount of linguistic processing is combined with a non-trivial topic model of documents.

4 Multinomial PCA

This sections introduces our topic model of documents from [3]. This is the multinomial PCA model, which has various other versions in the literature. Methods and models include non-negative matrix factorization [12] (which is a Poisson version), probabilistic latent semantic analysis [8], latent Dirichlet allocation [2], and generative aspect models [14]. A good discussion of the motivation for these techniques can be found in [8], a more sophisticated statistical analysis is by Minka [14]. We do not use Minka's methods here because while we have confirmed they produce higher-fidelity models, we have been unable to scale them appropriately for the large scale models we develop. MPCA in the large has a considerably different behaviour than for the small scale results reported by most authors (see [4]).

The bag of words model is as follows. Ignoring sparsity (which for the subsequent theory is an implementation issue), with J different words the i-th document becomes a vector $\boldsymbol{x}_i \in \mathcal{Z}^J$ where $x_{i,j}$ is the number of occurrences of the j-th word in the i-th document, and where the total $\sum_j x_{i,j}$ gives words in the document. There are I documents in total. Note that the bag-of-words representation is poor when it comes to dealing with the multiple senses of individual words (e.g., the use of "hot" in "hot coffee" and "hot car"), and is thus not very effective when performing semantic analysis of text, but it is a respectable first approximation.

4.1 The Model

A Gaussian model Consider Tipping *et al.*'s representation of standard PCA. A hidden variable $\boldsymbol{m}_i$ is sampled for the i-th document from K-dimensional Gaussian noise. Each entry represents the strength of the corresponding component and can be positive or negative. This is folded with the $K \times J$ matrix of component means $\boldsymbol{\Omega}$ to yield a J-dimensional document mean. Thus for the i-th document from the collection of I in total:

$$\begin{aligned} \boldsymbol{m}_i &\sim \mathit{Gaussian}(0, \mathbf{I_K}) \\ \boldsymbol{x}_i &\sim \mathit{Gaussian}(\boldsymbol{m}_i\boldsymbol{\Omega} + \boldsymbol{\mu}, \mathbf{I_J}\sigma) \end{aligned}$$

This relies on the data being somewhat Gaussian, which fails badly for document data where low and zero counts are normal. A square root transform can make the Poisson-like count data more Gaussian, but it still fails to treat zeros well.

The discrete analogue Modify the above as follows. First sample a probability vector $\boldsymbol{m}_i$ that represents the proportional weighting of components, and then

mix it with a matrix $\boldsymbol{\Omega}$ whose rows represent a word probability vector for a component:

$$\begin{aligned} \boldsymbol{m}_i &\sim Dirichlet(\boldsymbol{\alpha}) \\ \boldsymbol{x}_i &\sim Multinomial(\boldsymbol{m}_i \boldsymbol{\Omega}, L_i) \end{aligned}$$

where L_i is the total number of words in the i-th document, and $\boldsymbol{\alpha}$ is a vector of K-dimensional parameters to the Dirichlet. Thus the probability of the j-th word appearing in the i-th document is a convex combination of the document's hidden vector $\boldsymbol{m}_i$ and the j-th column of $\boldsymbol{\Omega}$. MPCA thus represents an additive/convex mixture of probability vectors.

Note also that a J-dimensional multinomial results from taking J independent Poissons and assuming their total count is given, thus a Poisson or multinomial analysis are similar and we just pursue the one here.

4.2 The Extremes of MPCA

With each document a vector $\boldsymbol{x} \in \mathcal{Z}^J$, traditional clustering becomes the problem of forming a mapping $\mathcal{Z}^J \mapsto \{1, \ldots, K\}$, where K is the number of clusters, whereas dimensionality reduction forms a mapping $\mathcal{Z}^J \mapsto \mathcal{R}^K$ where K is considerably less than J. Instead, MPCA represents the document as a convex combination. This forms a mapping $\mathcal{Z}^J \mapsto \mathcal{C}^K$ where $\mathcal{C}^K$ denotes the subspace of $\mathcal{R}^K$ where every entry is non-negative and the entries sum to 1 ($\boldsymbol{m} \in \mathcal{C}^K$ implies $0 \leq m_k \leq 1$ and $\sum_k m_k = 1$).

Suppose $\boldsymbol{m} \in \mathcal{C}^K$ is the reduction of a particular document. For multi-faceted clustering, $\boldsymbol{m}$ should have most entries zero, and only a few entries significantly depart from zero. For dimensionality reduction, $\boldsymbol{m}$ should have many non-zero entries and many significantly different from zero so that the reduced space $\mathcal{C}^K$ is richly filled out to make the dimensionality reduction efficient in its use of the K dimensions. A measure we shall use for this is entropy, $H(\boldsymbol{m}) = \sum_j m_j \log 1/m_j$. Thus multi-faceted clustering prefers low entropy reductions in $\mathcal{C}^K$ whereas dimensionality reduction prefers high entropy reductions. In the limit, when the average entropy is 0, the mapping becomes equivalent to standard clustering.

5 Theory Review

Here we summarise relevant probability modelling from [3]. A notation convention used here is that indices i, j, k in sums and products always range over 1 to I, J, K respectively, where i denotes a sample index, j a dictionary word index, and k a component index. I is the number of documents, J the number of words (dictionary size), and K the number of components. The index i is usually dropped for brevity (it is shared by all document-wise parameters) and is inserted when needed using the notation "$[i]$".

5.1 A Probability Model

Priors In the MPCA model $\boldsymbol{m}$, a K-dimensional probability vector, represents the proportion of components in a particular document. $\boldsymbol{m}$ must therefore represent quite a wide range of values with a mean representing the general frequency of components in the sample. A prior for such a probability vector is the Dirichlet, $\boldsymbol{m} \sim \text{Dirichlet}(\boldsymbol{\alpha})$. This is analytically attractive but otherwise has little to recommend it. The matrix $\boldsymbol{\Omega}$ represents word frequency probability vectors row-wise. A suitable prior for these is a Dirichlet whose mean correspond to the empirical frequencies of words occurring in the full data set, $\boldsymbol{f}$, so that $\boldsymbol{\Omega}_{k,\cdot} \sim \text{Dirichlet}(2\boldsymbol{f})$ (i.e., an empirical prior).

Likelihoods We briefly present the case where the bag of words has no order. The leads to identical theory to the so-called unigram model, or 0-th order Markov model which retains order. A hidden variable $\boldsymbol{w}$ is used which is a sparse matrix whose entry $w_{k,j}$ is the count of the number of times the j-th word occurs in the document representing the k-th component. Its row total $r_k = \sum_j w_{k,j}$ is the count of the number of words in the document representing the k-th component. Its column total $c_j = \sum_k w_{k,j}$ is the observed data. Denote by $\boldsymbol{w}_{k,\cdot}$ the k-th row vector.

$$\begin{aligned} \boldsymbol{m} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{r} &\sim \text{Multinomial}(\boldsymbol{m}, L) \\ \boldsymbol{w}_{k,\cdot} &\sim \text{Multinomial}(\boldsymbol{\Omega}_{k,\cdot}, r_k) \qquad \text{for } k = 1, \ldots, K \end{aligned}$$

The hidden variables here are $\boldsymbol{m}$ and $\boldsymbol{w}$ and the row and column totals are derived. The full likelihood for a single document $p(\boldsymbol{m}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\Omega})$ then simplifies to:

$$\frac{1}{Z_D(\boldsymbol{\alpha})} C^L_{w_{1,1}\ldots,w_{K,1},\ldots,w_{K,J}} \prod_k m_k^{\alpha_k - 1} \prod_{k,j} m_k^{w_{k,j}} \Omega_{k,j}^{w_{k,j}} \ , \tag{1}$$

where $Z_D()$ is the normalising constant for a Dirichlet. This is the likelihood used in constructing a variational EM algorithm with the hidden variables. It corresponds to one big multinomial because $\sum_{k,j} m_k \Omega_{k,j} = 1$. Thus the hidden variable $\boldsymbol{w}$ can be marginalized out to yield the formulation in Section 4.

$$\frac{1}{Z_D(\boldsymbol{\alpha})} C^L_{c_1\ldots,c_J} \prod_k m_k^{\alpha_k - 1} \prod_j \left(\sum_j m_k \Omega_{k,j} \right)^{c_j} \ .$$

6 Language Models for Search

The fundamental idea here is that a query is somehow generated by a probabilistic model based on a document. To test whether a document $\boldsymbol{x}$ matches a query, we turn the task around and compute the probability that the query matches the document. This goes as follows, for query $\boldsymbol{q}$ consisting of word

indexes $q_1, q_2, \ldots, q_Q$, we computer the probability for the query given the particular document to match with $\boldsymbol{x}$ and the broader context provided by the MPCA model for the document collection $\boldsymbol{\alpha}, \boldsymbol{\Omega}$. This is broken down with an independence assumption, as in [11].

$$\log p(\boldsymbol{q} \text{ matches} \,|\, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\Omega}) \;=\; \sum_{l=1,\ldots,Q} \log p(q_l \text{ matches} \,|\, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\Omega})$$

The match for an individual word is based on the notion that each word is either an exact match with one in the document, or a match with a word that could reasonably be generated by the model for the document assigned by MPCA. This distinction is maintained by a hidden indicator variable δ_l with value one for an exact match and zero for a possible match. It has a binomial distribution with parameter ρ. Thus

$$\begin{aligned}\log p(q_l \text{ matches}, \delta_l \,|\, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\Omega}, \rho) \;&=\; \rho^{\delta_l}(1-\rho)^{1-\delta_l} \\ &(\delta_l \, p(q_l, \,|\, \boldsymbol{x}, \text{ observed frequency}) + (1-\delta_l)\, p(q_l, \,|\, \boldsymbol{x}, \text{ MPCA model}, \boldsymbol{\alpha}, \boldsymbol{\Omega}))\end{aligned}$$

In this formulation, ρ is a free parameter and is estimated from the document-query combination. Thus

$$\begin{aligned}&\log p(q_l \text{ matches} \,|\, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\Omega}, \rho) \;= \\ &\quad \rho \, p(q_l, \,|\, \boldsymbol{x}, \text{ observed frequency}) + (1-\rho)\, p(q_l, \,|\, \boldsymbol{x}, \text{ MPCA model}, \boldsymbol{\alpha}, \boldsymbol{\Omega})\end{aligned}$$

where $p(q_l, \,|\, \boldsymbol{x}, \text{ observed frequency})$ is the observed frequency of the word q_l in the document and $p(q_l, \,|\, \boldsymbol{x}, \text{ MPCA model}, \boldsymbol{\alpha}, \boldsymbol{\Omega})$ is its probability according to the MPCA model for the document $\boldsymbol{x}$. The mixing parameter ρ is set to its average for the query on this document, which is approximately the proportion of exact matches for the query to the document and we return the final score using the approximation

$$p(q_l \text{ matches} \,|\, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\Omega}) \;\cong\; p(q_l \text{ matches} \,|\, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\Omega}, \hat{\rho})$$

7 Experiments

Our choice of document collection is the new Reuters Corpus[1], which contains 806,791 news items from 1996 and 1997. While there are a number of sizeable web collections available, especially from TREC, and we have our own large Amazon crawls, we chose to use the Reuters collection for these first experiments. Separately, we have been very pleased with the performance of the system on the Amazon crawl (and in fact demonstrate comparisons with Google to industrial partners on this crawl), the Reuters Corpus is a much better controlled baseline. It has few of the noise problems associated with the internet, and is much easier for us to understand, and thus interpret results. Because we want to evaluate

[1] Volume 1: English Language, 1996-08-20 to 1997-08-19.

long 100 word queries as well as shorter ones, the news environment is ideal since many of us can act as experts. Moreover, titles for these documents is usually a good indication of content, thus quick evaluation is feasible.

For our experiments we collected bag-of-words data from the new Reuters Corpus. The average length of a news item is 225 words, which translates to a total of 180 million word instances. About half a million of these are distinct words after the suffix "'s" is eliminated. We use simple statistical tests to identify compounds, together with a compound dictionary extracted from WordNet. We then use a standard fast tagger/lemmatizer system off the internet to represent words to a basic lexeme. Most commonly, this reduces plurals and possessive case. We kept the most frequent $65,000$ lexemes so generated. For instance, "Bill Clinton, Bill Cosby, Bill Gates" are compounds, but "Chelsea Clinton" is not.

Our code is written in C and C++ using Open Source tools and libraries such as the GNU Scientific Library (GSL). The GSL comes with a wide variety of distributional sampling algorithms as well as functions such as the digamma function and its derivatives. More details of some of the techniques used to scale up the MPCA algorithm appear in [4], and to our knowledge we are the only group applying this class of methods to over 50,000 documents at once. The new language models for querying have been implemented in a brute force manner doing a full pass of all documents, and thus is very slow at present. As a comparison, we compared our approach against the standard TF-IDF metric for comparing documents. The TF-IDF implementation uses the inverted index so is fast. Both algorithms throw away query words that are in the most 100 frequent words in the collection (i.e., approximately the stop words).

We evaluated a number of short queries, and a number of long queries. Long queries are text fragments, perhaps 100 words, of interest. These are usually not evaluated in search comparisons, however, we feel they will be a fundamental part of future search systems, computational difficulties not withstanding. The long queries were general news items taken off the internet on 12/6/03 and 13/06/03. These are given below:

1. Iraq war Spain
2. France tourism
3. Aero-engines manufacturer Rolls Royce
4. In a major discovery that may fill in the missing piece connecting us to our most immediate ancestors, the fossilized skulls of two adults and a child who lived in Ethiopia 160000 years ago could represent the earliest known remains.
5. Gregory Peck always seemed to be fighting for good causes. He did so in his most memorable performance as lawyer Atticus Finch in the 1962 film To Kill A Mockingbird and in dozens of other movies from his debut in 1944's Days Of Glory (as a Russian guerrilla) to his last (on TV) in 1993's The Portrait. Last week the American Film Institute named his role in To Kill A Mockingbird as the greatest movie hero of all time. It won Peck his only Oscar although he received four other nominations.

6. On the eve of an expected U.N. vote on the new global criminal court, human rights groups accused the Bush administration of using unconscionable tactics to undermine the tribunal rather than prosecute mass murderers in the 21st century.

Titles of the top twenty articles extracted are presented in the appendix. Evaluation of a short query by our new methods takes a few seconds on our 1.5GHz Pentium Linux machine, and the long queries takes several minutes. For the long queries, our new method is clearly superior, with TF-IDF performing poorly. Our new method is marginally superior in two of the short queries.

8 Conclusion

This is clearly preliminary work in the following sense:

- No attempt has been made to optimize or scale the search algorithm.
- Proper account of difficulties and methods relevant to the internet has not been made, for instance the use of hyper-link information, and linguistic preprocessing to overcome noise in internet text.
- Experiments need to be performed on true Internet data and with a greater variety of state-of-the-art search metrics.

However, the results are surprisingly good. In the larger queries, a service we believe will become critical in intranet contexts (and in the internet context if it could be scaled), our new methods returns results quiet acceptable to users. The quality of TF-IDF in the longer queries could not form the basis of a commercial system in our view. So we see tremendous opportunities here for developing intranet search applications where these kinds of methods should be computationally cost-effective.

Acknowledgements

This work was supported by the Academy of Finland.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
[2] D. Blei, A.Y. Ng, and M. Jordan. Latent Dirichlet allocation. In *NIPS*14*, 2002.
[3] W. Buntine. Variational extensions to EM and multinomial PCA. In *ECML 2002*, 2002.
[4] W.L. Buntine and S.Perttu. Is multinomial pca multi-faceted clustering or dimensionality reduction? In C.M. Bishop and B.J. Frey, editors, *Proc. of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.
[5] G. Forman. Choose your words carefully: An empirical study of feature selection metrics for text classification. In *PKDD 2002*, 2002.

[6] Moises Goldszmidt and Mehran Sahami. A probabilistic approach to full-text document clustering. Technical Report ITAD-433-MS-98-044, SRI International, 1998.
[7] K. Hall and T. Hofmann. Learning curved multinomial subfamilies for natural language processing and information retrieval. In *ICML 2000*, 2000.
[8] T. Hofmann. Probabilistic latent semantic indexing. In *Research and Development in Information Retrieval*, pages 50–57, 1999.
[9] T. Hofmann. Learning the similarity of documents. In *Advances in Neural Information Processing Systems 12*, pages 914–920, 2000.
[10] G. Karypis and E.-H. Han. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. In *CIKM-2000*, 2000.
[11] J. Lafferty and C. Zhai. Document language models, query models and risk minimization for information retrieval. In *SIGIR 2001*, New York, 2001.
[12] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.
[13] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.
[14] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *UAI-2002*, Edmonton, 2002.
[15] D. Moldovan, S. Harabagiu, R. Girju, P Morarescu, N.A. Lacatusu, A. Badulescu, and O. Bolohan. Lcc tools for question answering. In *Proc. 11th TREC*, 2002.
[16] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275–281, 1998.

9 Appendix: Results

In this appendix we list the titles of the top 20 documents selected. The new method is on the left, and the TF-IDF results are on the right.

9.1 Iraq war Spain

SPAIN: Iraq to use Syrian port to import aid - Rasheed.
SPAIN: PRESS DIGEST - Spain - Sept 12.
SPAIN: PRESS DIGEST - SPAIN - SEPT 13.
SPAIN: PRESS DIGEST - Spain - Sept 5.
IRAQ: PRESS DIGEST - Iraq - Nov 16.
UK: Iraq asks Spain to help Uday Hussein – opposition.
SPAIN: Spain says wishes U.S. had held off on Iraq attack.
IRAQ: PRESS DIGEST - Iraq - Feb 15.
SPAIN: Repsol to finalise Iraq oilfield deal - Rasheed.
IRAQ: PRESS DIGEST - Iraq - Dec 23.
FRANCE: European allies split over U.S. strike on Iraq.
SPAIN: PRESS DIGEST - Spain - Sept 4.
IRAQ: PRESS DIGEST - Iraq - July 22.
UK: PRESS DIGEST - London-based Arab newspapers - Jan 15.
IRAQ: Saddam's son telephones Iraq soccer team.
SPAIN: Iraq is fully cooperating with U.N. - Rasheed.
UK: Reuters Historical Calendar - September 13.
SPAIN: Spain to resume diplomatic activity in Iraq.
FRANCE: French spy photos put query on US Iraq move - daily.
REPUBLIC OF IRELAND: EU differs over U.S. raid on Iraq - Spring.

FRANCE: European allies split over U.S. strike on Iraq.
USA: Response to U.S. move shows anti-Iraq pact weaker.
SPAIN: PRESS DIGEST - SPAIN - SEPT 13.
REPUBLIC OF IRELAND: EU differs over U.S. raid on Iraq - Spring.
REPUBLIC OF IRELAND: EU differs over U.S. raid on Iraq - Spring.
SPAIN: PRESS DIGEST - Spain - Sept 12.
UK: Reuters Historical Calendar - September 13.
UK: Reuters historical calendar - August 23.
UK: Reuters Historical Calendar - February 15.
INDIA: Oil producers, consumers begin gathering at Goa meet.
UK: Reuters historical calendar - April 14.
UK: REUTERS HISTORICAL CALENDAR - FEBRUARY 22.
UK: UK's Labour promises defence review but no cuts.
UK: Reuters historical calendar - August 18.
UK: Reuters historical calendar - December 6.
FRANCE: France, Britain at odds over Zaire at summit.
UK: Reuters historical calendar - June 7.
UK: Reuters historical calendar - September 25.
UK: Reuters historical calendar - May 30.
UK: Reuters historical calendar - July 24.

Note the historical calendars retrieved by TF-IDF list important dates in the British Empire going back to the 1800's when Britian controlled Iraq. The

Spanish and Iraqi press digests list diplomatic and other exchanges between the countries relating to war and U.N. issues.

9.2 France tourism

FRANCE: Club Med sells stakes in Valtur, Situr.
MOROCCO: Moroccan tourism picks up in first seven months.
FRANCE: FRANCE-TELEVISION-MEDIAS-CANAL-ABONNEMENTS.
AUSTRALIA: Australia, Canada and French TV bodies in pact.
TUNISIA: Tunisian expatriates remit 754 mln dinars in 1996.
FRANCE: Algeria's Zeralda tourism firm loses 17 mln dinars.
MALTA: Malta tourism down 5.6 percent.
MALI: Impoverished Mali plugs into the Internet.
FRANCE: CYCLING-TOUR DE FRANCE TO START FROM IRELAND IN 1998.
FRANCE: GAUMONT-RESULTATS.
AUSTRALIA: RTRS-BRL Hardy joins with French wine-maker.
AUSTRALIA: TENNIS-RESULTS OF SYDNEY INTERNATIONAL TENNIS TOURNAMENT.
SPAIN: France most visited country in 1996 - study.
SPAIN: France most visited destination in 1996 - study.
FRANCE: Pathe buys TV channel from Landmark.
UK: International tourism seen growing through yr 2000.
BELGIUM: International anti-drugs swoop nets 123 people.
FRANCE: France to crack down on child sex.
SOUTH AFRICA: France bids to boost S.Africa sales, big projects.
TUNISIA: PRESS DIGEST - Tunisia - Jan 18.
CUBA: Cuba's tourism arrivals, income grow in 1996.
GERMANY: World tourism fair opens with environment worries.
MALTA: Malta tourism down 5.6 percent.
SPAIN: France most visited country in 1996 - study.
MOROCCO: Moroccan tourism picks up in first seven months.
SPAIN: France most visited destination in 1996 - study.
UK: International tourism seen growing through yr 2000.
CUBA: Despite U.S. embargo, Cuban tourism climbs.
MALTA: Malta tourism dips but appears to be recovering.
FRANCE: MEPs demand urgent clampdown on sex tourism.
SYRIA: 2.4 million tourists visited Syria in 1996.
ROMANIA: ROMANIA TO PRIVATISE HOTELS TO REJUVENATE TOURISM.
ROMANIA: Romania to privatise hotels to rejuvenate tourism.
SOUTH AFRICA: France bids to boost S.Africa sales, big projects.
MALTA: Malta tourism drops 5.3 percent in February.
MAURITIUS: Mauritius tourist arrivals up 17 pct in Q1 1997.
USA: International tourists to U.S. at all-time high.
FRANCE: Commission drafts measures to fight child sex tourism.
FRANCE: CYCLING-TOUR DE FRANCE TO START FROM IRELAND IN 1998.
FRANCE: EU prepares measures to fight child sex tourism.

9.3 Aero-engines manufacturer Rolls Royce

UK: Rolls Royce soars on aero engines hopes.
UK: Rolls Royce admits engine failures - paper.
UK: OPTIONS -Straddle sold in Rolls Royce.
UK: Rolls Royce says to sell Bristol Aerospace.
UK: OPTIONS - Straddles traded in British Telecom.
UK: OPTIONS - Straddles traded in Abbey National.
UK: OPTIONS - Volatility trades struck in UK equities.
UK: OPTIONS-B.A.T puts, calls sold for premium.
UK: OPTIONS - Short position rolled forward in BP.
UK: Air Canada to fly aero-engine to Boeing.
UK: Rolls Royce signs China training agreement.
UK: PRESS DIGEST - Financial Times - Sept 2.
UK: FOCUS - Britain commits to Eurofighter at air show.
FRANCE: Aero-engine firms see more exclusive sale pacts.
FRANCE: Monarch orders four more Airbus craft.
SINGAPORE: Rolls-Royce to raise engine profile in Asia.
UK: Rolls Royce wins $55 mln BA order.
UK: OPTIONS -CALL SPREAD SOLD IN ICI.
UK: Rolls Royce gets $500 mln Trent 700 order.
FRANCE: FOCUS-AMR to buy Brazil jets in $1 bln order.
UK: Rolls Royce soars on aero engines hopes.
UK: Rolls Royce says to sell Bristol Aerospace.
UK: OPTIONS -Straddle sold in Rolls Royce.
UK: Rolls Royce admits engine failures - paper.
UK: OPTIONS - Straddles traded in Abbey National.
UK: OPTIONS - Volatility trades struck in UK equities.
UK: OPTIONS - Straddles traded in British Telecom.
UK: OPTIONS-UK August stock options busy in thin trade.
UK: OPTIONS-B.A.T puts, calls sold for premium.
UK: OPTIONS - Short position rolled forward in BP.
UK: PRESS DIGEST - Financial Times - Sept 2.
UK: Rolls Royce signs China training agreement.
UK: Air Canada to fly aero-engine to Boeing.
UK: OPTIONS -CALL SPREAD SOLD IN ICI.
UK: FOCUS - Britain commits to Eurofighter at air show.
SINGAPORE: Rolls-Royce to raise engine profile in Asia.
FRANCE: Aero-engine firms see more exclusive sale pacts.
UK: Rolls says close to sale of steam power units.
UK: Rolls-Royce shares up as dividend raised.
UK: Steam power hits Rolls-Royce year profits.

9.4 Major discovery

In a major discovery that may fill in the missing piece connecting us to our most immediate ancestors, the fossilized skulls of two adults and a child who lived in Ethiopia 160000 years ago could represent the earliest known remains.

UK: We may be older than we think, scientist says.
UK: Oldest-ever stone tools found in Africa.
CANADA: FEATURE-Pow wow a return to Canadian Indian tradition.
RWANDA: FEATURE - School is Rwanda's grisly memorial to genocide.
SPAIN: Spanish scientists find possible new human species.
BANGLADESH: Bangladeshi men castrated by jealous wives.
BANGLADESH: FEATURE - Bangladesh struggles to end flesh trade.
UK: UK researchers find biological link to anorexia.
UK: Scientists trace 9,000-yr-old skeleton's kin.
MOZAMBIQUE: Thieves leave Mozambican airport in the dark.
SIERRA LEONE: Slave's song brings African American grandma home.
USA: Bones may be of Florida teens missing 18 years.
AUSTRALIA: RTRS-TIMELINES-Today in History - Dec 24.
UK: Aboriginal tells Britain to return ancestor's head.
CHAD: FEATURE - Blind Chadian teenage pianist is symbol of hope.
USA: Brazil insect helps protect African cassava crops.
GREECE: FEATURE - Philip of Macedon's tomb reveals secrets.
AUSTRALIA: YEAREND - Misfits with grudges were year's worst killers.
USA: Funeral for six-year-old strangled in Colorado home.
BHUTAN: FEATURE - Sleepy Bhutan awakens to tourism.
SOUTH AFRICA: Revered skull of S.Africa king is Scottish woman's.
EGYPT: Egypt experts dig up cancer in ancient skull.
BELGIUM: Police sound final whistle on skull kick-about.
ETHIOPIA: Donors pledge $2.5 billion for Ethiopia.
EGYPT: Lonely Egypt widow digs ups dead husband's skull.
ETHIOPIA: Ethiopia economy set to grow, central banker says.
ETHIOPIA: Ethiopia farm production not affected by drought.
ETHIOPIA: Ethiopia seeks to renegotiate Russian loans.
JAPAN: Whales, hippos, cows share ancestor - Japan research.
AUSTRALIA: RTRS-Broker crosses 10 pct of Discovery.
ETHIOPIA: Ethiopia says financial sector underdeveloped.
UK: Scientists push back date of earliest humans' era.
AUSTRALIA: RTRS-Discovery says forms new oil alliance.
ETHIOPIA: Ethiopia achieved record cereal harvest in 1995.
ETHIOPIA: Ethiopia says sold $1 bln at auctions in 96/97.
ETHIOPIA: Ethiopia names new agriculture minister.
ETHIOPIA: Ethiopia appeals for $2 billion in foreign aid.
ETHIOPIA: Ethiopia's justice minister resigns.
AUSTRALIA: RTRS-Premier plans talks with Discovery.
AUSTRALIA: RTRS - Discovery rejects Premier bid.

9.5 Gregory Peck

Gregory Peck always seemed to be fighting for good causes. He did so in his most memorable performance as lawyer Atticus Finch in the 1962 film To Kill A Mockingbird and in dozens of other movies from his debut in 1944's Days Of Glory (as a Russian guerrilla) to his last (on TV) in 1993's The Portrait. Last week the American Film Institute named his role in To Kill A Mockingbird as the greatest movie hero of all time. It won Peck his only Oscar although he received four other nominations.

ITALY: Venice Film Festival embraces art in all forms.
ITALY: Venice Film Festival embraces art in all forms.
USA: Actress Marjorie Reynolds dies aged 77.
USA: "The English Patient" takes lion's share of Oscars.
USA: Films in Robert Mitchum's career.
CZECH REPUBLIC: Czech Oscar winners to make film about Britain's RAF.
FRANCE: Star-studded gala, Wenders film mark Cannes 50th.
FRANCE: France revels in Binoche's surprise Oscar win.
FRANCE: Star-studded gala, Wenders film mark Cannes' 50th.
USA: Maine library tracks celebrity reading.
ITALY: Venice prize goes to IRA film "Michael Collins".
FRANCE: French film, theatre actress Casares dies.
MEXICO: France honours "rebel beauty" of Mexico screen goddess.
CZECH REPUBLIC: CZECHS GIVE OSCAR WINNERS BEER TOAST IN HOMECOMING.
USA: Legendary director Fred Zinnemann dies at 89.
CANADA: Quebec movie director and actress die in plane crash.
UK: Wilde the new Oscar winner in Britain.
FRANCE: French heap praise on Italian actor Mastroianni.
FRANCE: Egypt director preaches tolerance after film ban.
FRANCE: France revels in Binoche's win.
UK: Briton sues over televised suicide bid.
USA: American Mobile Satellite names CFO.
USA: Cyberian Outpost shuts real doors as online sales soar.
NETHERLANDS: INTERVIEW-POLYGRAM SAYS FILM PLAN ON TRACK.
NETHERLANDS: INTERVIEW-PolyGram says film plan on track.
USA: N.Y. official sees '97 film, TV production gains.
USA: Peck Comm. Schools, Mich., Aa - Moody's.
TAIWAN: China director, movie sweep Taiwan's "Oscars".
INDONESIA: Indonesia's film hopes pinned on blockbuster.
USA: Oldest complete U.S. movie found in Oregon basement.
USA: Hollywood's independent studios not so independent.
USA: Hollywood's independent studios not so independent.
INDONESIA: FEATURE-Indonesian film industry hopes pinned on blockbuster.
USA: Peck Comm Sch Dist, Mich. won by Wm. Hough.
USA: FEATURE-Hollywood writers lament sorry state of film.
USA: Oscar nominations for best foreign film.
GERMANY: German film feted as opening act of Berlin fest.
UK: Film director Zinnemann dies of heart attack.
GERMANY: Global film buyers look for hits at Berlin market.
GERMANY: Global film buyers look for hits at Berlin market.

9.6 New global criminal court

On the eve of an expected U.N. vote on the new global criminal court, human rights groups accused the Bush administration of using unconscionable tactics

to undermine the tribunal rather than prosecute mass murderers in the 21st century.

UNITED NATIONS: U.S. urges U.N. assembly action on human rights.
UNITED NATIONS: U.S. INSISTS ON U.N. COUNCIL ROLE IN CRIMINAL COURT.
UNITED NATIONS: Main issues in establishing world criminal court.
USA: U.S. human rights group criticizes major powers.
AUSTRIA: Austria calls for human rights action plan.
SWITZERLAND: US criticises Chinese tactics at UN rights forum.
USA: Clinton puts political spin on human rights speech.
UNITED NATIONS: UN rebukes Burma for continued political supression.
UNITED NATIONS: Albright leads U.N. criticism of Burma government.
USA: Group charges powers with human rights hypocrisy.
SWITZERLAND: World jurists body says Tunisia seized lawyer.
VATICAN: Vatican demands West help end world hunger.
switzerland: Amnesty slams UN on China, Turkey, Algeria abuses.
SWITZERLAND: RIGHTS-CHINA-DILEMMA (NEWS ANALYSIS, SCHEDULED).
UNITED NATIONS: Clinton signs test ban treaty, urges further steps.
ITALY: Gaddafi sees more Islamic attacks in U.S. - paper.
CAMBODIA: Cambodia's new drug law comes under fire.
UK: Monthly Review - Big refugee exodus from Zaire.
MALI: Christopher campaigns for democracy on Africa tour.
NORWAY: China dissident Wei nominated for 1997 Peace Prize.
UNITED NATIONS: Prospects grow for an international criminal court.
NETHERLANDS: Funds row forces U.S. lawyers out of U.N. tribunal.
NETHERLANDS: Plans for world criminal court held up by haggling.
USA: VarTech acquires 21st Century Professional.
TANZANIA: U.N. Rwanda tribunal head defends his record.
UNITED NATIONS: U.N. moves to clean up Rwanda tribunal structures.
TANZANIA: U.N. Rwanda genocide tribunal receives key accused.
RWANDA: Rwanda says U.N. genocide tribunal useless.
TANZANIA: UN investigates Rwanda genocide tribunal.
SOUTH KOREA: S.Korea seeks open market in 21st century.
KENYA: Rwanda genocide tribunal employee murdered.
NETHERLANDS: U.N. tribunal soldiers on despite Bosnia inertia.
NETHERLANDS: U.N. tribunal judges criticise Bosnia conference.
SWITZERLAND: Swiss court rules Rwandan may face U.N. tribunal.
SWITZERLAND: US criticises Chinese tactics at UN rights forum.
YUGOSLAVIA: Yugoslavia lets UN war-crimes tribunal open office.
SWITZERLAND: Amnesty slams U.N. human rights forum.
TANZANIA: Defence says genocide tribunal has no jurisdiction.
NETHERLANDS: War crimes tribunal on ex-Yugoslavia - key facts.
NETHERLANDS: Albright to visit U.N. war crimes tribunal.