Representing Interests as a Hyperlinked Document Collection

Michelle Fisher
University of Exeter
Exeter, EX4 4QF, UK
M.J.Fisher@exeter.ac.uk

Richard Everson
University of Exeter
Exeter, EX4 4QF, UK
R.M.Everson@exeter.ac.uk

ABSTRACT

We describe a latent variable model for representing a user's interests as a hyperlinked document collection. By collecting hyper-text documents that a user views, creates or updates whilst at their computer, we are able to use not only the content of these documents but also the inter-connectivity of the collection to model the user's interests. The model uses Probabilistic Latent Semantic Analysis and Probabilistic Hypertext Induced Topic Selection and decomposes the user's document collection into a set of factors each of which represents a user's interest. This model can be used to personalise information access tasks such as a personalised search engine or a personalised news service. Our latent variable model's performance is compared with that of a more conventional vector space clustering algorithm.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.3.4 Information Storage and Retrieval: Systems and Software – *user profiles and alert services*

General Terms: Algorithms, Experimentation.

Keywords: User interests, hyperlinked/hypertext document collections, latent variable models, information access.

1. INTRODUCTION

In this age of information it is common knowledge that the information we require, be it a web page, new research papers on a certain subject or simply the answer to a question, cannot be accessed easily. The tens of thousands of results from a search engine must be tackled or the research paper repositories must be searched; we spend far too much time accessing the information rather than using it. A solution to these problems is personalised information access. A personalised search engine, in which queries the user gives are automatically enhanced according to the user's interests and the results are ranked in order of relevance to the user's information need, a paper tracker which discovers new papers which are relevant to the user's interests and a personalised electronic newspaper that tracks the user's changing interests are all potential applications of personalised information access.

The usual method of representing a user's interests for information personalisation is by means of a *profile* [15]. Traditionally profiles are represented by a small list of weighted keywords that are manually entered and updated by the user. There are many problems with this method. With a limited number of terms, term matching methods are not satisfactory; for example, there may be two documents describing the same topic, but the terms used in the two documents are different, using a term matching method the similarity of these two documents could not be detected. Another example of term matching inadequacies are homonyms; many terms have several meanings, but term matching methods only have one representation for each term so the different meanings of a single term cannot be detected. A second problem is that the user has to enter and update the keywords that represent their interests. The problem here is that the user will seldom have the time, inclination or skill to think of terms that represent their interests. A third problem is that users cannot describe their interests in such a way that the resulting profile can be used to produce successful personalised information access because a good term matching profile will need an exhaustive list of terms that describe each interest area. A good example here is that a typical search engine query consists of only one or two terms [16] which is insufficient to describe a detailed information need. It is unlikely users will treat a profile any differently.

We aim to show in this paper that profiles can automatically be built by monitoring the hyper-text documents that a user views and creates whilst at their computer. There have been previous attempts to automatically build user profiles, but they always rely on user feedback to understand the interests [14]. Soltysiak and Crabtree [15, 4] also tried to build keyword profiles by monitoring the web pages and e-mails users viewed; in section 3 we compare the representation of user interests gained from their conventional vector space clustering algorithm to our latent variable model.

In this paper we present a model for representing the user's interests as a hyper-linked document collection. The documents we monitor from the user's activity at their machine can tell us about the user's interests in two ways: the first is the content of the documents; the second is the citations or links between those documents. Previously, only the terms or content in documents have been used to describe user's

CIKM'03 November 3–8, 2003, New Orleans, Louisiana, USA.

interests and, while there is much information to be gained from the terms in the documents, the links between these documents also provide important additional information.

A document generally contains citations or references to other documents that expand upon the topics covered or help explain points in the original document. Link analysis studies (e.g. [11]) have shown that analysis of the link structure within a document collection can separate subject areas (documents generally cite documents from the same subject area), and find the most important documents on these subject areas (documents which are cited by many documents in the same subject areas).

Hence, if the user's interests can be represented as a document collection we expect that the links within the collection can be used to separate each of the user's interests and find the important documents in each of these interests.

Traditionally, link analysis methods were applied only to references made in research papers to discover whether there was overlap between subjects, who the most important authors were, what the most important papers in a subject area were, etc. Recently, however, link analysis methods have been applied to the WWW, where the citations are now called hyperlinks. The popular web search engine Google [8] has successfully used WWW link information to enhance the retrieval and ranking of search results [13, 1]. Here we apply link analysis methods to a new area, namely personalisation, where we define citations or hyper-links simply as 'links between documents'. To do this we expand the notion of a 'document' and a link or citation between documents.

We regard the user as a document (although with unobservable content). Then, when the user views a web page they are making a citation to the web page; the web page, in turn, may contain citations or links to other documents such as other web pages (via hyperlinks) or people (via e-mail addresses). When the user sends an e-mail she is making a citation both to the e-mail document and also to the people who are to receive the e-mail. The e-mail document itself contains citations (via message-ids) to the recipients of the e-mail and to the e-mail documents mentioned in the 'References' and 'In-Reply-To' headers. Modelling all of these types of links within the user's document collection enables us to better understand the user's interests.

We use a latent variable model to represent the user's interests, similar to the Probabilistic LSA and Probabilistic HITS model presented by Cohn and Hofmann [9, 2, 3]. Projecting the term and citation count vectors in to a lower dimensional space, the latent semantic space, furnishes a succinct, smoothed representation of the user's interests. This is because documents in the user's document collection with generally co-occurring terms and citations will have a similar representation in latent semantic space even if they have no terms or citations in common.

In this paper we present a static view of the user's interests, but current work involves extending this model to reflect the user's changing interests using either sliding time windows or a hidden Markov model version of the model presented here.

We first describe and discuss the hyperlinked model. Section 3 discusses experiments illustrating the model's properties and gives a comparison between our latent space model and a vector space model. Section 4 discusses the results and future work.

2. A HYPERLINKED MODEL

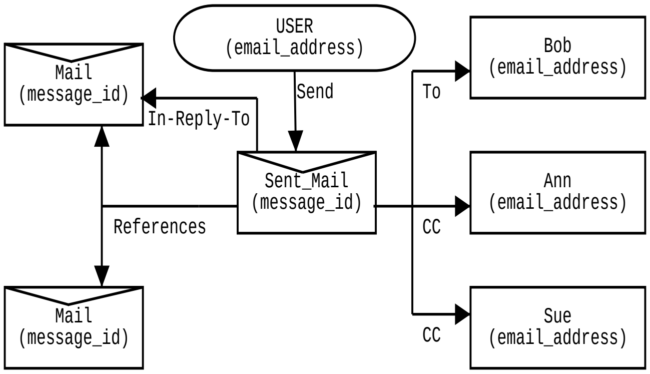

Figure 1: The links created when the USER sends an e-mail.

We monitor all textual documents that the user views or creates whilst at their computers. This includes web pages they view on the WWW, e-mails they send, e-mails they receive, Usenet news they view and any files they create or update on their machine. We will refer to this set of documents as the *USER's* document collection.

These documents comprise both content and links. In common with most information access systems, we represent the content as a bag-of-words, each term carrying a weight reflecting its probability of occurrence. As shown in [3] and [7], both content and links are important in discovering the latent topics from this hyper-linked document collection. Using the content and links of documents we may imagine creating a 'user interest space' in which all documents that the user has seen lie. Documents pertaining to similar subjects are close to each other, whereas documents from areas with nothing in common lie far apart. Both content and links shape this space. The overlap in the subject matter of the content of documents from the same subject area means that content helps shape the space and, as discussed in the introduction, citations from documents are generally to documents from a similar subject area so links are also useful.

In our model we view the person being monitored as the *USER* and all other people (identified by their e-mail addresses) contained in the model as documents. People documents can have in-links, created when the *USER* sends an e-mail to the person or when the content of a document cites a person (see figure 1). People documents can also have out-links, created when the person sends an e-mail to the *USER* or someone else with the *USER* CC'ed. However, the content of the people documents are unobservable variables.

Usenet newsgroups are represented in a similar manner to people documents. Newsgroup documents can contain in-links when they are cited from the content of another document or when someone sends an e-mail to that newsgroup. But the contents of newsgroup with relation to the *USER's* interests are unobservable variables.

It is perhaps strange to think of documents with links, but unobservable content. However, it is analogous to having a document with content but no links, where the links can be thought of as being unobservable variables. Using the generative model described in 2.1, we could infer the probable content of the people or newsgroup documents with relation to our *USER* and also infer probable links of documents containing unobservable links.

Each of the document types has an identifier: web pages have URLs; e-mails and Usenet news have message-ids; people have e-mail addresses; and files have path names. In our model, we regard the mention of one of these identifiers in a document as citing or linking to the document. Web pages and people are the only document types to have global identifiers, that is, these are the only documents that can be cited from any document content. The message-ids of e-mails and news articles can be used to find links between e-mails and news but not other document types.

The content of all types of documents can contain citations to web pages via URLs, to people via e-mail addresses and to newsgroups via Uniform Resource Identifiers (URIs).

Different types of documents have different link and content qualities, we will discuss each document type in turn.

Web Pages Viewed by USER. When the *USER* views a web page, they are making a direct citation to that document. The web page document content may contain citations to people documents via e-mail addresses, to other web pages via URLs and to newsgroups via URIs.

Files Created or Updated by USER. When the *USER* creates or updates files they are making a direct citation of them. The files may also contain citations in the form of URLs, URIs or email addresses within the content of the document.

E-mail Sent by the USER *(Figure 1).* E-mails and news articles contain meta-data in the form of headers, the meta-data can be parsed to find links between e-mails and news. E-mails are directed documents like telephone calls or postal mail. When a user sends an email they direct it to another user or users, whereas when a user views a web page, news article or file that document is static. The *USER* creates sent e-mails and files rather than simply viewing them, future work involves reflecting this difference by giving additional weight to these important documents. When the user sends an email, they are directly citing the contents of that e-mail document and, because the *USER* is *sending* the e-mail, they are also making direct links to those people who are to receive the e-mail. E-mails can be sent to one or more recipients, but can only have one sender. For example in figure 1, the one sender is *USER* and there are three recipients Bob, Sue and Ann. From the e-mail document itself there are citations to the recipients (discovered from the 'To', 'CC' and 'BCC' headers).

As well as the In-Reply-To header, which allows an e-mail to refer to the message to which it replies, the e-mail standard RFC822 provides for an e-mail to refer to (possibly more than one) e-mail in a thread of e-mail exchanges via the 'References' header. All the documents/e-mails mentioned in the 'References' and 'In-Reply-To' headers are also included as citations. Also, although not shown in figure 1, the sent e-mail may also contain links to web pages and people within the body of the e-mail.

E-mail Received by the USER. In the case that Bob sends the *USER* an e-mail, many of the same rules apply. Again there can be only one sender and one or more recipients. From the received e-mail there are citations to the recipients of the e-mail and to any messages mentioned in the 'In-Reply-To' and 'References' headers. Again, the received e-mail may cite web pages and people within the body of the e-mail.

Usenet News Articles Read by the USER. When the *USER* views a Usenet news article they make a link to that news article, but not to the other (unknown) recipients of the article, because they are only viewing the article and not sending it. Like normal e-mails, news articles can have only one sender, but one or more recipients with the first recipient always being the newsgroup. We regard the news article itself as citing the sender and the recipients. It also cites any messages in the 'In-Reply-To' and 'References' headers. Like other documents, the news article may also have citations within its content.

Each document in the *USER's* document collection is thus represented by a set of terms (with term counts) together with a set of citations (out-links) and citation counts. In the next section we show how a latent variable model, based on PLSI and PHITS [3], may be used to extract the principal elements of the *USER's* interests.

2.1 PLSI & PHITS: A Model for Content and Links

PLSI & PHITS [3] is a latent variable model, in which the high dimensional term and citation data is projected onto a smaller number of latent dimensions. This results in noise reduction, topic identification and is a principled method of combining text and link information. PLSI and PHITS are probabilistic equivalents, appropriate for multinomial observations, of the LSI [5] and HITS [11] methods.

In common with the majority of information retrieval methods, we ignore the order of the terms and citations within a document, and describe a document $d_j \in \mathcal{D} = \{d_1, ..., d_J\}$ as a bag-of-words or terms $t_i \in \mathcal{T} = \{t_1, ..., t_I\}$ and citations or links $c_l \in \mathcal{C} = \{c_1, ..., c_L\}$. This information can also be described by a term-document matrix N, where entry $N_{j,i}$ contains the number of times term t_i occurs in document d_j, and a document-citation matrix A, where entry $A_{j,l}$ corresponds to the number of times citation c_l occurs in document d_j.

PLSI & PHITS [3] is a latent variable model for general co-occurrence data, which associates an unobserved class variable $z_k \in \mathcal{Z} = \{z_1, ..., z_K\}$ with each observation or occurrence of term t_i or citation c_l in document d_j. The model is based on the assumption of an underlying document generation process:

- Pick a document d_j with probability $P(d_j) = 1/J$
- Pick a latent class, or *interest*, z_k with probability $P(z_k|d_j)$
- Generate a term t_i with probability $P(t_i|z_k)$ and a citation c_l with probability $P(c_l|z_k)$.

The observed document consists of observation pairs (d_j, t_i) and (d_j, c_l), but the latent class variable z_k is discarded and is not observed. Note however, that the terms and citations occurring in a particular document are associated because they are each conditioned on the particular latent class z_k. Thus terms and links occurring in a particular document are expected to be associated with particular topics associated with the document.

As shown by Cohn and Hofmann [3], the joint probability model for predicting citations and terms in documents can be expressed as:

$$P(d_j, t_i) = \sum_k^K P(z_k)P(t_i|z_k)P(d_j|z_k) \quad (1)$$

$$P(d_j, c_l) = \sum_k^K P(z_k)P(c_l|z_k)P(d_j|z_k) \quad (2)$$

$P(t_i|z_k)$, $P(c_l|z_k)$, $P(d_j|z_k)$ and $P(z_k)$ are determined by maximising the normalised log-likelihood function of the observed term and citation frequencies. Contributions from term information and citation information are combined as a convex combination:

$$\begin{aligned} \mathcal{L} \;=\; & \alpha \sum_j \sum_i N_{j,i} \log P(d_j, t_i) \\ + \; & (1-\alpha) \sum_j \sum_l A_{j,l} \log P(d_j, c_l) \end{aligned} \quad (3)$$

The parameter α sets the relative weight of terms and link information. If α is 1 the model takes only the terms into consideration, while if it is 0 only citations are considered.

The Expectation Maximisation (EM) [6] algorithm, a standard method for maximum likelihood estimation in latent variable models, can be applied to find the local maximum of $\mathcal{L}$.

2.2 Applying the Model

The low-dimensional latent space representation of the user's interests, may be applied to information access tasks such as personalised search engines, newspapers and paper trackers as shown in sections 3.4 and 3.5. These tasks are either information filtering (IF) problems, such as a personalised newspaper where new documents (news articles in this case) arrive and it must be determined whether they are of interest to the *USER*, or information retrieval (IR) problems in which documents retrieved in response to a query must be ranked in order of how well they match the *USER*'s interests. The *USER*'s profile may also be used to enhance the original query. Both IR and IF need a measure of how similar a document or query is to other documents in the *USER*'s interests. We calculate the similarity in K-dimensional latent space as follows. First, a representation of the new document or query, q, is found by projecting or folding q in by calculating mixing proportions by EM iteration during which the factors are fixed and only the mixing proportions $P(z_k|q)$ are calculated in each M-step.

The cosine similarity in latent space can then be used to discover how similar the query document q is to each of the documents in the *USER*'s collection.

$$sim(d_j, q) \;=\; \frac{\sum_k P(z_k|q)P(z_k|d_j)}{\sqrt{\sum_k P(z_k|q)^2}\sqrt{\sum_k P(z_k|d_j)^2}} \quad (4)$$

3. RESULTS

The web pages, e-mails and news articles that a *USER* creates or views are collected by intercepting the client-server protocols (HTTP, SMTP, POP and NNTP), files that are created or updated by the *USER* are also collected. Document pre-processing steps are taken to transform these documents into lists of terms and citations with frequency counts.

3.1 USER Document Collections

We present results from documents collected from four users; a post-doctoral researcher, a PhD student, a departmental secretary and a lecturer all in the department of computer science at the University of Exeter from August 2002 until December 2002. Table 1 shows the numbers of documents used for each user for these experiments.

postdoc	phd	secretary	lecturer
4866	4366	4555	4760

Table 1: Document collection sizes for each user.

A brief description of each of the users to give insight into the results is as follows.

postdoc Post-doctoral researcher, he states his research interests as: neural networks, evolutionary computation, optimisation, data structures and Markov chain Monte Carlo reversible jump methods, with personal interests in the martial art of Tae Kwon Do and current affairs.

phd PhD student, her interests are information retrieval, statistical pattern recognition, the python programming language and popular science.

secretary Departmental secretary, her work related interests are student problems, admission queries, meeting arrangements, and local and national companies who may be of interest to students searching for jobs. Her personal interests are news about the Exeter and Durham areas, books and DVDs.

lecturer Lecturer, his research interests include: Pattern recognition and independent component analysis; Bayesian MCMC methods for signal and image processing; multiobjective optimisation and machine learning methods for global optimisation.

3.2 Experimental Details

In the results reported here we have used the user's web pages, Usenet news articles, sent e-mails and received e-mails; however, we have not used (the relatively small number of) files created or updated.

Full text indexing was used for all documents in the *USER* document collections. All terms were stemmed using Porter's stemming algorithm [12] and stop-words were removed. After stop-word removal the 1500 most frequently occurring terms were used.

For all of the experiments shown here we used random initial starting values for $P(z_k)$ and $P(z_k|d_j)$ and Tempered EM [10] with 20% held out data, $\eta = 0.95$ and a lower computational temperature limit of 0.7.

3.3 Representation of Interests

We begin by illustrating how the latent space factors provide a representation of the user's interests.

Table 2 shows a selection of the factors for **secretary** obtained with 32 factors and $\alpha = 0.7$. We have found empirically that $0.6 \leq \alpha \leq 0.8$ is most effective in distinguishing between distinct user interests; $0.6 \leq \alpha \leq 0.8$ gives most, but not all, weight to term information: a detailed study of the effects of α on text classification can be found in [7]. Although not shown here, we have investigated profiles with

Factor1	Factor2	Factor3	Factor4
Text			
univers	confer	school	room
exet	comput	friend	reserv
depart	research	add	seminar
scienc	recognit	find	resourc
comput	neural	secondari	tutori
contact	network	contact	pleas
email	pattern	member	februari
school	imag	club	august
undergradu	languag	year	januari
student	analysi	work	novemb
staff	model	leav	decemb
Citations			
Uni of Exeter homepage	Exeter DCS Homepage	Friends Reunited	Exeter DCS reserving resources
Exeter e-mail address	Exeter DCS MSC course	Friends Reunited	Exeter DCS reserving resources
Exeter e-mail address	Exeter DCS MSC modules	Friends Reunited	Exeter DCS reserving resources
Exeter e-mail address	Exeter DCS MSC books	Friends Reunited	Exeter DCS reserving resources
Exeter DCS homepage	Exeter DCS MSC projects	Friends Reunited	Exeter DCS reserving resources
Exeter e-mail address	Exeter DCS MSC news	Friends Reunited	Exeter DCS reserving resources
Exeter DCS people	Exeter DCS AIIE group	Friends Reunited	Exeter DCS reserving resources
International office uni of Exeter	Exeter DCS lecturer	Friends Reunited	Exeter DCS reserving resources

Table 2: secretary: The highest probability terms and links from some PLSI & PHITS factors. 'Exeter e-mail address' indicates addresses of staff and students at the University of Exeter; 'DCS' is department of computer science.

different numbers of latent factors K; we found that increasing K increased the level of detail found in the factors.

Factor1 appears to be about the department of computer science at the University of Exeter with assorted links to University of Exeter and departmental web pages and to the e-mail addresses of members of staff in the department. Factor2 represents the **secretary**'s involvement in the master's course in autonomous systems which has courses on pattern recognition, neural networks etc. Factor3 shows the **secretary**'s recent interest in the popular Friend's Reunited website. Finally, Factor4 is interesting because it represents a more specific aspect of her work: here the probable terms are months of the year and the links are mainly to the department of computer science's calendar for reserving resources such as seminar rooms and laptop computers.

We also compared our model to Crabtree and Soltysiak's conventional vector space clustering model [4] where each document is represented as a vector of term weights. The weight of a term i in a document j is $w_{i,j} = (\log(tf_{i,j} + 1))/(\log df_i)$ where $tf_{i,j}$ is the number of times term i occurs in document j and df_i is the number of documents in which term i occurs. Details of the clustering algorithm can be found in [4]. To repeat Crabtree and Soltysiak's work, we sum the $w_{i,j}$ scores of all documents in each cluster and the 12 most highly scoring terms are used to represent that cluster. We have found that Crabtree and Soltysiak's method produces clusters with highly weighted terms that are indicative of the user's interests. However, for all four users over half of their documents are not assigned to a particular cluster – it is unlikely that this proportion of their documents were outliers. Also, again for all users, about 20% of the documents were clustered into one generic cluster which could not be identified as an interest. The rest of the clusters were small, generally consisting of about 10–20 documents. The clusters shown in tables 4 and 5 are examples of these small clusters. The clusters were chosen because they represent the same interests as those shown in tables 2 and 3.

Using the PLSI + PHITS factors there is no one factor which has a significantly higher probability than the others and all documents are used in the model. We emphasise that the clustering method only considers term information whereas PLSI + PHITS considers both term and links. Also, a document can belong to only one cluster using the vector space clustering, whereas PLSI + PHITS can represent multi-topic documents.

Cluster1	Cluster2	Cluster3	Cluster4
comput	modul	happygroup	room
scienc	system	friend	resourc
messag	introduct	limit	reserv
am	com	friendsreunit	tutori
cours	msc	reunion	seminar
studi	comput	board	pleas
subject	student	opinion	august
forward	autonom	find	projector
admiss	inform	messag	laptop
exet	neural	repres	februari
univers	coursework	great	th
dear	cours	club	novemb

Table 4: secretary: Terms with largest weights from some clusters using Crabtree and Soltysiak's method.

Lecturer's PLSI & PHITS factors are shown in table 3 using 64 factors and $\alpha = 0.7$. The first factor, Factor1, clearly shows **lecturer**'s research interest in pattern recognition and more specifically Gaussian mixture models, the articles at Citeseer were also concerning mixture models. Factor2 represents **lecturer**'s involvement in the cognitive science students' research projects. Factor3 is very interesting because

Factor1	Factor2	Factor3	Factor4
Text			
model	scienc	backup	test
mixture	cognit	tape	arrai
cluster	credit	linux	numer
data	research	disk	integ
statist	comput	system	return
gaussian	psycholog	configur	shape
estim	project	server	matrix
classif	student	network	scipi
likelihood	modul	compress	doubl
EM	dept	unix	dot
space	essai	comput	linalg
fit	supervisor	softwar	vector
Citations			
Google Search MCMC	Exeter e-mail address	Amanda homepage	Delphi Numeric
Google Search MCMC	Exeter e-mail address	Sourceforge Amanda	Message ID
Google Search Statistics	Exeter e-mail address	Information on Amanda	Message ID
Google Search MCMC	Exeter e-mail address	Amanda guide	MathFIT
Citeseer Paper	Message ID	Message ID	Delphi FAQ
NZ Uni. Statistics Dept.	Message ID	Tape prices	Message ID
Citeseer Paper	Uni of Exeter DCS	Sourceforge Amanda	Message ID
BTexact Agents	Uni of Exeter DCS	Sourceforge Amanda	NumPy e-mail address

Table 3: lecturer: The highest probability terms and links from some PLSI & PHITS factors. 'Exeter e-mail address' indicates addresses of staff and students at the University of Exeter; 'DCS' is department of computer science; 'MCMC' stands for Markov Chain Monte Carlo; Citeseer paper refers to results for an article at CiteSeer NEC research index site; NumPy is a library for matrix computations using the python programming language.

Cluster1	Cluster2	Cluster3	Cluster4
inform	com	tape	arrai
signal	modul	gb	numpi
distribut	comput	drive	rank
equat	skill	backup	discuss
pattern	studi	di	list
understand	level	code	net
structur	scienc	product	python
recognit	program	id	listinfo
contain	essai	price	sourceforg
nois	intellig	softwar	thinkgeek
uai	method	order	http
oxfor	student	request	sf

Table 5: lecturer: Terms with largest weights from some clusters using Crabtree and Soltysiak's method.

it represents a temporary interest, at the time **lecturer** was investigating backup systems for linux computers, in particular the Amanda software as is evident from both the terms and the links. The final interest, factor4, represents the **lecturer**'s interest in matrix computation and linear algebra.

Table 5 shows the clusters for **lecturer** using Crabtree and Soltysiak's clustering algorithm, again we chose these clusters because they are closest to those in table 3.

Other interesting features that appear in all four of the PLSI + PHITS profiles. We frequently obtain 'stop-word' factors that contain terms that could be considered stop-words but are not in the standard stop-word list; for example, *do, get, go, re, just, love, work, am, like, think, want, know, thing, realli*. Also common are 'site overview' factors, where the most probable links are all to one site and the most probable terms appear to have come from the home page of that site, for example, terms: *yahoo, uk, car, ireland, new, sport, financ, shop, person, employ, search, job, pick, centre, tv* and links: *Yahoo, Yahoo mail, Yahoo directory, Yahoo education, Yahoo news, Yahoo directory, Yahoo, Yahoo search*. There are also 'formatting' factors in which all terms relate to the formatting of different document types and the citations are often advertising web pages or web pages containing a lot of complicated formatting that was not removed whilst parsing, for example: *font, color, famili, px, size, arial, serif, text, weight, helvetica, san, decor, bold, verdana, td*. We anticipate that deleting the factors with low information content will facilitate the filtering of extraneous noise from the representations of profiles.

3.4 Query Expansion

The previous section showed some of the factors representing interests obtained using the latent variable model, this section shows how these representations can be applied to personalise an information access task, namely query expansion.

Having folded in the query, an augmented query is composed by adding the most probable terms and citations from the most probable factor, $z^* = argmax(P(z_k|q))$.

Alternatively, a smoothed representation $P(t_i|q)$ and $P(c_l|q)$ may be obtained by back projection. The query is then expanded using the most probable terms and citations from $P(t_i|q)$ and $P(c_l|q)$.

As an example, we used the ambiguous query 'python' (is the user interested in snakes, Monty Python's Flying Circus, the programming language, etc?). Using the PLSI + PHITS

Terms			
1. numer	2. sourc	3. code	4. modul
5. librari	6. plot	7. linux	8. gener
Citations			
1. SciPy homepage		4. SciPy tutorial	
2. SciPy FAQ		5. Sourceforge Numeric	
3. SciPy module			

Table 6: phd: The most probable terms and citations in the most probable factor z^* given the query 'python'.

Terms			
1. python	2. numer	3. sourc	4. code
5. modul	6. librari	7. plot	8. packag
Citations			
1. SciPy homepage		4. Sourceforge Numeric	
2. SciPy tutorial		5. Sourceforge homepage	
3. SciPy information			

Table 7: phd: The most probable terms and citations given the query ('python'), $P(t_i|q)$ and $P(c_l|q)$.

model we can return useful results for ambiguous queries by first expanding them with terms and citations and presenting the expanded query to a conventional search engine. Most conventional search engines allow the user to use URLs in their queries, pages that are similar to the web page identified by the URL or web pages that link to the URL are then returned in the user's search results.

We used **phd**'s profile with $K = 32$ and $\alpha = 0.7$, she had recently been working with Python's Numeric libraries for matrix computations and SciPy, scientific tools for Python. Tables 6 and 7 show expansion terms and links from the two expansion methods. The expansion terms and citations clearly represent her current interest in scientific tools and the python language. Augmenting a query to a conventional search engine with these terms and citations returns results more relevant to **phd**'s interests.

As a second illustration we used **postdoc**'s profile with 32 factors and $\alpha = 0.8$. **postdoc** is a Tae Kwon Do instructor so we used the query 'martial art' for which a conventional search engine would return results concerning all martial arts. Tables 8 9 augmenting terms and citations for the two expansion method; both methods have produced query terms and citations that would direct the rather general search 'martial art' toward **postdoc**'s interest in Tae Kwon Do.

These results demonstrate that the PLSI & PHITS model can be used to enhance queries with regard to the *USER*'s interests.

3.5 Ranking Search Results

Documents retrieved from a conventional (non-personalised) search engine can be ranked in the context of the *USER*'s interests in the following two ways.

Firstly, for each search result document we can evaluate the likelihood of that document, d_j, belonging to the *USER*'s document collection using:

$$\mathcal{L}_j = \alpha \sum_i N_{j,i} \log P(d_j, t_i) + (1 - \alpha) \sum_l A_{j,l} \log P(d_j, c_l) \tag{5}$$

Terms			
1. sport	2. citi	3. british	4. do
5. council	6. tae	7. kwon	8. com
Citations			
1. Exeter Tae Kwon Do		4. Amateur martial assoc.	
2. Amateur martial assoc.		5. Message ID	
3. Amateur martial assoc.			

Table 8: postdoc: The most probable terms and citations in the most probable factor z^* given the query 'martial art'.

Terms			
1. sport	2. citi	3. martial	4. art
5. tae	6. council	7. kwon	8. do
Citations			
1. Exeter Tae Kwon Do		4. Amateur martial assoc.	
2. Youth Hostels		5. Amateur martial assoc.	
3. Message ID			

Table 9: postdoc: The most probable terms and citations given the query ('martial art'), $P(t_i|q)$ and $P(c_l|q)$.

The search results can be ranked in order of the likelihood that they belong to the *USER*'s document collection.

Alternatively, documents can be ranked by folding in both the query q and the document d_j retrieved from the conventional search engine, and the cosine similarity measure used to obtain a score of how relevant the new document d_j is to the *USER* using equation (4). The search result documents can then be ranked in order of their relevance score.

One of **lecturer**'s interests is *independent components analysis* also known as ICA. Table 10 shows a conventional search engine's results for the highly ambiguous query 'ICA'; note that the first relevant pages are ranked 13^{th} and 14^{th}. We used **lecturer**'s profile and the un-enhanced query 'ICA' to rank the top 50 documents returned from a conventional search engine according to **lecturer**'s interests using the likelihood document ranking method. Table 11 shows the first 15 results returned from the personalised document ranking method. Ranking according to this method gives the first relevant documents at 2^{nd} and 3^{rd}. There were only 4 documents about *independent components analysis* found in the top 50 results from the conventional search engine and all of these are found in the top 12 results using the personalised method. The majority of the non-English results returned were found at the bottom of the ranked list using the personalised method.

4. CONCLUSIONS

In this paper we have shown that a user's interests may be represented by a latent space model of a hyper-linked document collection and have given preliminary results demonstrating that this model can be used to improve information retrieval tasks. We find that the inclusion of a quite general notion of hyper-links aids the discriminatory power of the latent space factors. In contrast to the majority of existing personalisation schemes (eg. [14]), these profiles are multi-dimensional, they are automatically learned and do not require direct specification by the user. The latent space factors provide a degree of interpretability.

Conventional Search Engine Results	Notes
1. Swedish: Food page	Non-English
2. International council on archives	
3. Institute of contemporary arts	
4. International communications association	
5. International co-operative alliance	
6. International cartographic association	
7. Lab. for computer comms and apps.	
8. New media centre	
9. Consultancy company	
10. Swedish magazine	Non-English
11. Web design and Internet provider	
12. (as 5)	
13. **Independent components analysis**	Relevant
14. **Independent components analysis**	Relevant
15. Spanish: Mexican site	Non-English

Table 10: Results from a conventional search engine for the query 'ICA'.

PLSI & PHITS results	Notes
1. Inter. commissions for acoustics	
2. **Independent components analysis**	Relevant (was 24th)
3. **Independent components analysis**	Relevant (was 13th)
4. Inter. council for IT	
5. Jigsaw puzzles	
6. **Independent components analysis**	Relevant (was 37th)
7. Japanese school of computing	Non-English
8. Computing consultancy company	
9. Crime prevention site	
10. (as 4)	
11. **Independent components analysis**	Relevant (was 14th)
12. Inter. cartographic association	
13. Institute of contemporary arts	
14. Computing company	
15. Inter. co-operative alliance	

Table 11: Ranked results from the PLSI & PHITS model for the query 'ICA' using lecturer's profile. Inter. stands for international.

There are a number of interesting avenues still to be explored. In this work we have considered only citations to documents in the user's document collection. Recent work [7] shows that using citations to documents external to the collection can be very beneficial for information access tasks. We anticipate that using these *external* documents will be effective for representing user interests. Likewise, we expect that 'stemming' citations (in a manner analogous to the way in which terms are stemmed) and using only the most frequently occurring citations will improve both the representation and the information retrieval performance of this model. It will also be of importance to determine the importance of different document types (sent email, received email, Usenet news, web pages and files) to the representation of the user's interests: we anticipate that documents that the user *creates*, rather than merely views, will be most important.

The profiles generated here are static. We are currently developing hidden Markov models for evolving profiles to represent changing interests.

The security of personalisation information is, of course, an important issue not addressed here. We emphasise that model presented here can be implemented entirely on the user's *local* computer – there is no need to expose private correspondence etc to untrusted, external search engines or information sources. Nonetheless, the latent space factors could provide a compact, effective format for the interchange of personalisation information.

Acknowledgments

Michelle Fisher is supported by a CASE studentship with BT and EPSRC. We would like to thank Gareth Jones for his valuable comments.

5. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[2] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proc. 17th International Conf. on Machine Learning*, 167–174, 2000.

[3] D. Cohn and T. Hofmann. The missing link – a probabilistic model of document content and hypertext connectivity In *Advances in Neural Information Processing Systems*, 13:430–436, 2001.

[4] B. Crabtree and S. Soltsiak. Identifying and tracking changing interests. In *Journal on Digital Libraries*, 2(1):38–53, 1998.

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. Am. Soc. Info. Sci.*, 6:391–407, 1990.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm with discussion. *J. Royal Statistical Soc., B*, 39:1–38, 1977.

[7] M. J. Fisher and R. M. Everson. When are links useful? experiments in text classification. In *Advances in IR, 25th European Conference on IR research, ECIR*, 41–56, 2003.

[8] Google. `google.com/technology/whyuse.html`.

[9] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd Annual International ACM SIGIR Conference on R. & D. in IR*, 50–57, 1999.

[10] T. Hofmann and J. Puzicha. Unsupervised learning from dyadic data. Technical Report TR-98-042, Berkeley, CA, 1998.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[12] M.F.Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Tech. report, Stanford Digital Library Technologies, 1998.

[14] M. Pazzani, J. Muramatsu, and D. Billsus. Syskill and Webert: Identifying interesting web sites. In *AAAI/IAAI*, 1:54–61, 1996.

[15] S. Soltysiak and I.B. Crabtree. Automatic learning of user profiles - towards the personalisation of agent services. *BT Technology Journal*, 16(3):110–117, 1998.

[16] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proc. of the 19th Annual International SIGIR Conf. on R. & D. in IR*, 4–11, 1996.