When are links useful?

Experiments in Text Classification.

Michelle Fisher and Richard Everson
M.J.Fisher@exeter.ac.uk, R.M.Everson@exeter.ac.uk

Department of Computer Science, Exeter University

Abstract. Link analysis methods have become popular for information access tasks, especially information retrieval, where the link information in a document collection is used to complement the traditionally used content information. However, there has been little firm evidence to confirm the utility of link information. We show that link information can be useful when the document collection has a sufficiently high link density and links are of sufficiently high quality. We report experiments on text classification of the Cora and WebKB data sets using Probabilistic Latent Semantic Analysis and Probabilistic Hypertext Induced Topic Selection. Comparison with manually assigned classes shows that link information enhances classification in data with sufficiently high link density, but is detrimental to performance at low link densities or if the quality of the links is degraded. We introduce a new frequency-based method for selecting the most useful citations from a document collection for use in the model.

1 Introduction

In recent years link analysis methods have become popular for use in information access tasks. Many of the popular search engines [15] now use link analysis to assist their ranking algorithms. The most well known of these search engines is Google [7]; Brin et al. describe the link analysis method (known as PageRank [14]) used by Google. Kleinberg described his link analysis method called Hypertext Induced Topic Selection (HITS) in [12]. However, until recently no-one had attempted to show that link methods were actually beneficial to information access task performance. In 1999 TREC [16] started a web track with one of the key aims being to discover whether link methods were useful. Unfortunately, the TREC-8 experiments showed that combining content and link information yielded no benefits over using pure content methods [9]; in fact the performance of some participants was actually degraded when link information was included. This was in part due to a poorly collected document collection containing few cross-host links, but the more recent TREC Web Tracks, with specially engineered document collections with high link density across hosts have still not shown that links are particularly useful [8]. There have been a couple of victories for link methods: Bailey, Craswell & Hawking showed that they could be useful for site finding tasks [2], and Craswell, Hawking & Robertson

also obtained good results for site finding tasks, using link anchor text information [4] rather than the links themselves. Use of link anchor text may be one of the reasons that search engines using link analysis methods, such as Google, seem to perform so well even though there is no documented performance benefit from combining term and link information.

Ideally we would have used information retrieval experiments to give the results in this paper, unfortunately there are very few adhoc information retrieval datasets available containing both text and links. The comparison of datasets shown in this paper would not have been possible with current information retrieval datasets, so instead, to give quantitative results, we have used text classification experiments.

This paper shows that link information can be useful when the document collection being used has a sufficiently high link density. The quality of the link information is also important. We use Cohn & Hofmann's Probabilistic LSI and Probabilistic HITS (PLSI and PHITS) method [3]; we repeat and extend Cohn & Hofmann's text classification experiments in [3] on the Cora [13] and WebKB [1] data sets, which are collections of automatically classified academic papers and manually classified web pages. Comparison with manually assigned classes shows that link information enhances classification when the link density is sufficiently high, but is detrimental to classification performance at low link densities or if the quality of the links is degraded. We also introduce a new frequency-based method for selecting the most useful citations from a document collection for use in the model; this has the added benefit of increasing the effective link density of a collection.

We first describe Cohn & Hofmann's joint probabilistic model for content and links, PLSI and PHITS, section 3 discusses link density measures and in section 4 we describe the document collections used in our experiments. The results of our experiments are presented and discussed in sections 5 and 6.

2 PLSI & PHITS

PLSI & PHITS [3] is a latent variable model, where the high dimensional 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 of the LSI [5] and HITS [12] methods appropriate for multinomial observations.

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 $c_l \in \mathcal{C} = \{c_1, ..., c_L\}$. Note that there may be $L > J$ documents available for citation; $L = J$ when citations are made entirely within the document collection. 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 doc-

ument 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.

PLSA & 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 and 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 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. 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 Hofmann [10], the joint probability model for predicting citations and terms in documents can be expressed as:

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

$P(t_i|z_k)$, $P(c_l|z_k)$, $P(d_j|z_k)$ 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:

$$\mathcal{L} = \alpha \sum_j \sum_i N_{j,i} \log P(d_j, t_i) + (1 - \alpha) \sum_j \sum_l A_{j,l} \log P(d_j, c_l) \quad (2)$$

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 $\mathcal{L}$. Alternating the E and M steps increases the likelihood, $\mathcal{L}$, converging to a local maximum. Hofmann [11] introduced tempered EM (TEM) to avoid over-fitting, which is often severe for sparse data, by controlling the effective model complexity and to reduce the sensitivity of EM to local maxima. TEM is discussed in some detail since our procedure differs from that of Cohn and Hofmann [3]. A control parameter β modifies the E-step by discounting [10] the data likelihood when $\beta < 1$; $\beta = 1.0$ results in the standard E-step. To implement TEM we use the algorithm proposed by Hofmann [10]: hold out some portion of the data, perform early stopping on the held out data with $\beta = 1$. Decrease β by setting $\beta = \eta\beta$, where $\eta < 1$, and continue TEM steps

while performance on held out data improves. While a lower limit of β has not been reached or until decreasing β does not yield further improvements, continue decreasing β and performing TEM steps.

Classification of an unknown test document d_{new} is achieved as follows. First, a representation in latent space $P(z_k|d_{new})$ is calculated by 'folding in'; that is, the mixing proportions $P(z_k|d_{new})$ are calculated by EM, but with the factors $P(t_i|z_k)$ and $P(c_l|z_k)$ fixed. The K-dimensional vector of mixing proportions is then used with the cosine similarity measure to find the nearest neighbour in the training documents to d_{new}. The test document is then assigned the class of its nearest neighbour, and classification accuracy is judged against the manually assigned class.

3 Link Density Measures

Document collections containing links (for example, a collection of journal articles where the links between documents are the citations, or a collection of documents from the web where the links are hyperlinks) can be modelled as a directed graph $\mathcal{G}(\mathcal{D},\mathcal{C})$. Here the vertices $d \in \mathcal{D}$ are the documents and each edge $c \in \mathcal{C}$ is a directed link between a pair of documents. The edges each have an integer weight w_c; this weight is 1 when a document cites another document only once, but can be higher when there are multiple citations from one document to another.

A measure of the link density of a document collection could be defined in several ways, but we find the graph sparseness, Γ, measure most useful:

$$\Gamma = \frac{|\mathcal{C}|}{J \cdot L} \quad \leq \quad \Gamma_w = \frac{1}{J \cdot L} \sum_{c=1}^{|\mathcal{C}|} w_c \tag{3}$$

where $|\mathcal{C}|$ is the number of edges in the graph, J is the number of vertices or documents and L is the number of documents that can be cited. Often J and L are equal but defining them separately will be useful in later sections. Note that Γ is the fraction of non-zero entries on the citation matrix, A. The weighted graph sparseness, Γ_w, measure accounts for the weights on the edges of the graph; that is, multiple citations carry additional weight.

A further useful measure is the average number of links, ρ, that a document makes and the weighted average number, ρ_w, of links per document:

$$\rho = \frac{|\mathcal{C}|}{J} \quad \leq \quad \rho_w = \frac{1}{J} \sum_{c=1}^{|\mathcal{C}|} w_c \tag{4}$$

4 Document Collections

The WebKB Collection The WebKB data set [1] contains 8282 web pages collected from computer science departments of various universities in January 1997 by

Table 1. Number of documents in each manually determined class for the WebKB dataset (left) and the Cora dataset (right)

Cora	
Class	*Number of Docs*
Reinforcement_Learning (ReinL)	354
Genetic_Algorithms (GA)	625
Theory	532
Probabilistic_Methods (ProbM)	656
Case_Based	491
Rule_Learning (RuleL)	282
Neural_Networks (NN)	1390
TOTAL	4330

WebKB	
Class	*Number of Docs*
course	907
department	176
faculty	1091
project	497
staff	129
student	1599
TOTAL	4399

the WebKB project of the CMU text learning group. The pages have been manually classified into seven classes. To permit direct comparison with Cohn & Hofmann's experiments, we ignored the seventh 'other' category containing 3764 of the documents and also ignored web pages with non-conforming HTML. The resulting collection contains 4399 web pages, each document belonging to one of six classes (table 1). For the WebKB set we use full text indexing. In total there are 30403 terms in the index, of which we use the 500 most frequent stemmed terms that do not occur in standard stop-words lists. We found (experiments not shown here) that increasing the number of terms used yielded little benefit in classification accuracy.

The WebKB data set contains 4395 within collection links, the graph sparseness link density is $\Gamma = 2.27 \times 10^{-4}$ and the average number of links per document is $\rho = 0.999$. The weighted link density is $\Gamma_w = 2.58 \times 10^{-4}$ and the weighted average number of links per document is $\rho_w = 1.13$.

To calculate the term density and average number of terms, the $\mathcal{C}$s are exchanged for $\mathcal{T}$s and the Ls for Is in equations (3) and (4). The term density is 9.9×10^{-2} and the weighted term density is 1.8×10^{-1}. The average number of terms per document is 49.7 and the weighted average is 92.4.

2020 of the documents in the WebKB collection are not cited by any document within the WebKB collection and 2793 documents do not cite any documents within the collection.

The Cora Collection The Cora data set [13] was collected automatically by intelligent web spiders. In the whole data set there are about 37000 papers, all of which have been automatically classified into hierarchical categories such as /Artificial_Intelligence/Machine_Learning/Theory. In common with Cohn & Hofmann's work, we use the 4330 documents in the 7 sub-categories of Machine Learning, see table 1.

Terms occurring in the title, abstract, author and affiliation are used as index terms. In total there are 15753 terms; after stop-word removal, documents are indexed by the 500 most frequently occurring stemmed terms.

The document collection contains 12263 within collection links, using the graph sparseness measure it has a link density of $\Gamma = 6.54 \times 10^{-4}$ and $\rho = 2.8$ links per document. Each paper in the Cora data set may make multiple citations of the same document throughout the text, so that the edges $\mathcal{G}(\mathcal{D}, \mathcal{C})$ have weights greater than one. However, we have used only the 'references' section of the papers, in which each document is only cited once. Consequently, all edges have weight one, meaning that the weighted link density measures are equal to the non-weighted ones.

The term density is 7.2×10^{-2}, the weighted density is 1.1×10^{-1}. The average number of terms per document is 36.1 and the weighted average is 55.8.

2115 of the documents are not cited from any of the documents within the Machine Learning collection and 993 documents do not cite any documents within the collection.

It is already clear from this brief collection analysis that the Cora dataset not only has a higher link density, but also has higher quality link information than the WebKB collection. Only just over one fifth of the Cora documents do not cite any documents within the collection, whereas almost two thirds of the documents in the WebKB document collection do not cite any documents within the dataset.

5 Experiments and Results

To investigate whether link information can be useful for information access tasks if there is a sufficiently high link density in the document collection, we use the PLSI and PHITS model to gain a low-dimensional latent space representation of the data sets and then perform nearest neighbour text classification on a 15 percent held out portion of the documents. We first performed experiments on both the WebKB and Cora data sets to serve as our baseline results and then altered various aspects of the data sets and models to explore how link density and link quality affect classification accuracy.

As described in section 2, the parameter α assigns more or less weight to the content and link information of the collection. When $\alpha = 0$ the model uses just link information and when $\alpha = 1$ only term information is used. Any value in between 0 and 1 will produce a joint representation using both term and link information. The experiments below show whether using both content and link information produces better classification accuracy than using either alone.

For the TEM regime in our experiments (introduced in section 2), we use an initial value of $\beta = 1$, a lower limit of $\beta = 0.8$ and an update parameter of $\eta = 0.95$. Cohn and Hofmann used an update parameter of $\eta = 0.9$ in [3] which caused the model to stop learning too early; however, our TEM parameters allow us to improve the classification accuracy on both document collections, but also show that best performance on the WebKB data set is achieved using only term information.

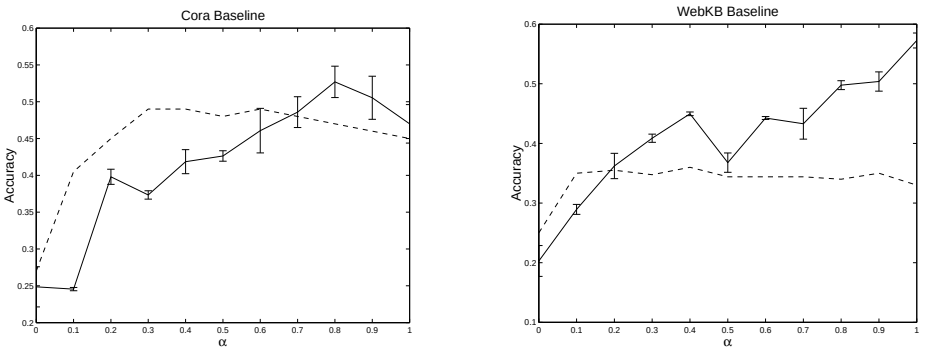

Fig. 1. Average classification accuracy and standard deviation for the Cora (left) and the WebKB (right) document collections. Dotted lines show Cohn & Hofmann's classification accuracies estimated from published graphs [3].

For these experiments we use the same number of factors as there are true classes, seven for Cora and six for WebKB. Each of the figures shows the average classification accuracy over three runs with different randomly selected test sets, where accuracy is the fraction of documents correctly classified.

Figure 1 shows the classification accuracy for the Cora and WebKB document collections as the weight, α, ascribed to link and term information is varied. Classification rates obtained by Cohn & Hofmann [3], using different TEM settings are also shown. In agreement with Cohn & Hofmann, we find that the Cora documents are more accurately classified when both link and term information is utilised, although with our TEM regime, peak accuracies are achieved when more weight is given to term information ($\alpha = 0.8$) than reported by Cohn & Hofmann. However, the addition of link information ($\alpha < 1$) in the WebKB data is detrimental to classification performance. It should be noted, however, that the TEM regime used here is able to achieve substantially higher classification rates for these data than previously reported [3].

As demonstrated below, the contrasting results for Cora and WebKB collections are due to the quality and density of the link information in both of these collections. There is an almost three times higher link density in the Cora data set than in the WebKB data set. Also, the quality of the link information (i.e., citations of other papers) in the Cora set is likely to be higher because they are academic papers. The WebKB data set was collected from only a few universities and many of the links in this collection are likely to be navigation links (links to help the user navigate around a web site). As discussed by Kleinberg [12], navigational links are less likely to point at relevant information; in fact, Kleinberg went as far as to delete navigational links in his experiments. Hawkings [8] suggests that the most important type of links in a web collection are links from one host to another, links between universities are rare in the WebKB data.

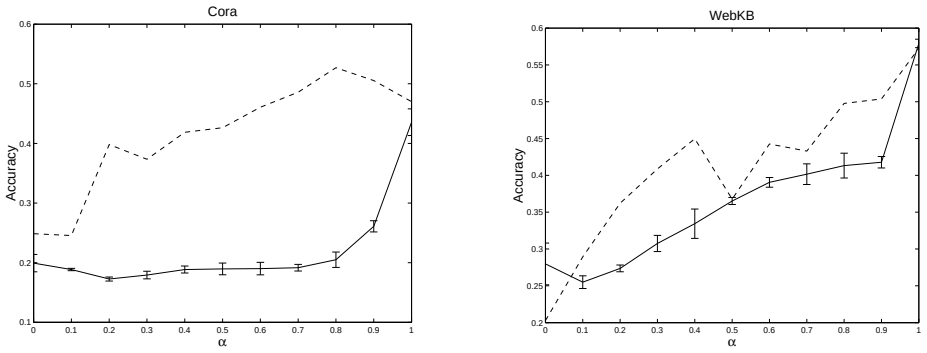

Fig. 2. Average classification accuracy using randomised link information for the Cora (left) and the WebKB (right) data sets. The baseline accuracies are shown by the dotted lines (see figure 1).

5.1 Randomising the Link Information

To determine whether the link information in both datasets provided useful information, we randomised the link information in both collections. In terms of the graph $\mathcal{G}(\mathcal{D}, \mathcal{C})$ described in section 3, we kept the number of edges and the weights on the edges the same, but simply made the edges point from and to different vertices or documents.

As shown in figure 2, the classification accuracy for the Cora collection drops considerably for any $\alpha < 1$. However, the classification accuracy for the WebKB data is barely changed; the $0 < \alpha < 1$ accuracies are slightly lower and the trend with α is more linear. This indicates that the link information in the WebKB collection is ineffectual for text classification.

5.2 Diluting Link Information

To show that the main reason for the poor classification accuracy when using link information for the WebKB document collection is the low link density, we have run experiments on the Cora document collection in which we removed portions of the link information. To reduce the amount of link information in the Cora data set to the density in the WebKB data set, we randomly delete two thirds of the edges in the Cora graph. Note that this procedure reduces the quantity of links but leaves the quality unchanged, whereas the randomisation procedure reduces the quality but leaves the link density constant. The results of this experiment are shown in figure 3. We carried out the same procedure a second time but this time removed only one third of the edges, resulting in the classification accuracies shown on the right hand side of figure 3.

When both one third and two thirds of the link information is removed there is no benefit to using both content and link information, supporting the hypothesis that low link density is the reason for poor accuracy in the WebKB collec-

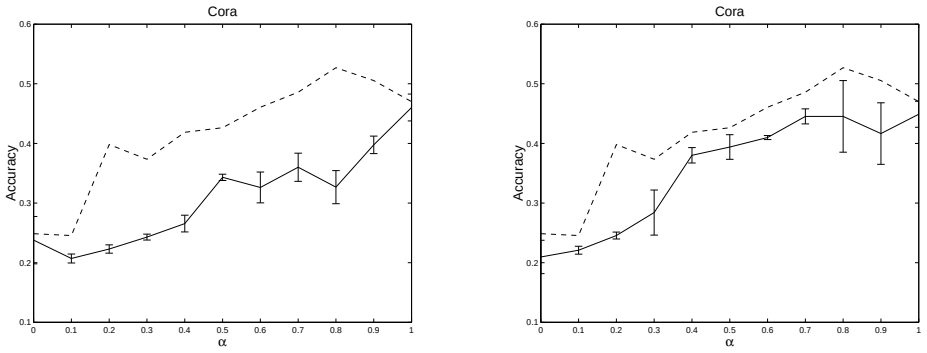

Fig. 3. Removing portions of the link information from the Cora dataset. Average classification accuracy and standard deviation when two thirds of all the link information is removed is shown on the left, ie. making Cora link densities equivalent to WebKB. The right shows the same information when only one third of the link information has been removed. The dotted lines denote the baseline accuracies (see figure 1).

tion, although, removing only one third results in classification rates that are only slightly below the baseline result. It is interesting to observe the resemblance between the classification rate curves for the diluted Cora data and the unaltered WebKB data (figure 1).

Also, notice that diluting the links even by two thirds results in better classification rates than when the link information was randomised (figure 2). This confirms that even small amounts of (high quality) link information are useful in the Cora data set and that misdirected links are detrimental to classification performance.

5.3 Concentrating Link Density

Although we have used an unsupervised training method, the classes of the documents are in fact known. The WebKB data has been manually classified and the Cora data was automatically classified. This class information can be used to add additional links into the collections in a manner that reinforces the true topic distributions. This is achieved by randomly choosing two documents from within the same true class and making an intra-class link between them; no additional inter-class links were made.

For the Cora data set, double the total number of original links were added in intra-class links, which has the effect of raising the intra-inter class link ratio from $9568/12263 = 0.78$ to $34094/36789 = 0.9$. The left hand side of figure 4 shows the classification accuracy achieved; clearly the accuracy has been raised over a wider range of α than in the baseline experiments, and the peak accuracy is now achieved by giving more weight (lower α) to the link component of the model.

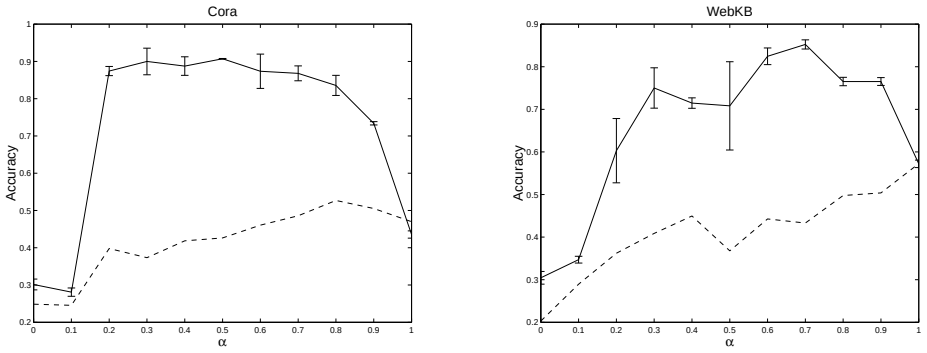

Fig. 4. The average classification accuracy and standard deviation with artificially increased link densities for the Cora (left) data and the WebKB (right). Baseline accuracies are shown by the dotted lines (see figure 1).

Initially, we increased the number of intra-class links in the WebKB data set by two thirds, as with Cora, to $9674/13185 = 0.71$. However, the classification accuracy using this higher link density did not improve. To discover the reason for this, we examined in more detail the intra-class to inter-class link ratio of both document sets and the quality of the link information in both of the datasets.

Tables 2 and 3 show the numbers of links between each class. The Cora data set has a intra-inter link ratio of $9568/12263 = 0.78$, and as can be seen from table 2 the majority of links are intra-class links. It may also be observed that the matrix is roughly symmetrical, meaning that documents in class A are cited by documents in class B about as many times as documents in class B cite documents in class A. The total number of links from each class is also roughly proportional to the number of documents in that class.

In contrast the WebKB data set has much lower quality link information: the intra-inter link ratio is $884/4395 = 0.201$, the links are not symmetrical and the number of links per class is not proportional to the number of documents in that class. It is interesting to note here that the documents in the 'department' class (only 179, mostly home page documents) contain very few out links (59) but are cited the most times out of all the classes (1648).

All of the documents in the Cora data set serve the same purpose: they are all journal papers about computer science. On the other hand, although all the of the web pages in the WebKB data set are about computer science departments, web pages in general have more diverse functions. A department home page, for example, is for introducing the surfer to and directing the surfer around a site, whereas a staff leaf node is simply for reading.

Taking into account the differences between the data sets, we increased the intra-inter link ratio of the WebKB data as well as concentrating the link density of the both data sets. We increased the link density until the intra-inter

Table 2. Number of links from and to each manual class for the Cora data. The largest number of out links from each class is shown in bold and the largest number of in links to each class is underlined.

	C Base	Rule L	Theory	Rein L	NN	GA	Prob M	Total Out
C Base	**924**	82	136	27	97	38	51	1365
Rule L	46	**470**	113	3	27	5	6	670
Theory	84	119	**1369**	47	181	38	140	1978
Rein L	40	12	55	**1522**	104	164	46	1943
NN	33	27	186	94	**2200**	54	174	2768
GA	20	2	31	62	56	**1724**	5	1900
Prob M	25	16	121	22	103	3	**1359**	1649
Total In	1172	728	2011	1777	2768	2026	1781	12263

Table 3. Number of links from and to each manual class for the WebKB data.

	course	department	faculty	project	staff	student	Total Out
course	**258**	143	175	22	4	121	723
department	1	**31**	5	20	2	0	59
faculty	132	**488**	176	176	4	63	1039
project	4	143	186	**194**	33	129	689
staff	0	**76**	14	34	18	3	145
student	214	**767**	330	213	9	207	1740
Total In	609	1648	886	659	70	523	4395

class link ratio was the same as that of the link-concentrated Cora data, namely $33185/36696 = 0.9$. The right hand side of figure 4 shows the resulting classification accuracy. Like the Cora data, the accuracy has been raised over a wide range of α, and improved classification now results from using link information in addition to term information.

To summarise: document classification is enhanced by using link information if the links are both sufficiently dense and of sufficiently high quality.

5.4 Number of Factors

Cohn & Hofmann's experiments [3] used the same number of factors or latent classes, K (eq (1)), as the number of manually assigned classes. We investigated whether the classification accuracy was improved by using more factors. As shown in figure 5 classification is improved for both datasets when the number of factors is doubled. However, further doubling the number of factors again showed no further improvement. This indicates that the collections should be represented by more topics than indicated by the manual categorisation; one possibility is that this effect is due to multi-topic documents aswell as single-topic documents on a wide variety of subjects. It is also worth noting that the peak classification rates are achieved at the same link:content ratio (α) as with the smaller number of classes.

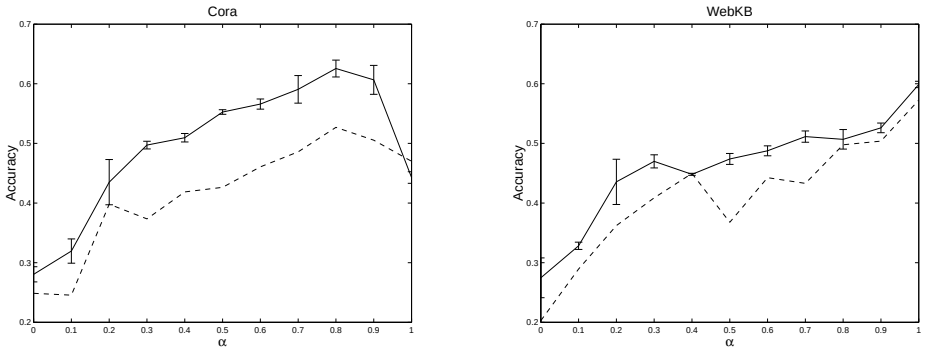

Fig. 5. Average classification accuracy for the Cora data (left) and the WekKB data (right) with 14 latent factors (Cora) and 12 latent classes (WebKB). The dotted lines show the baseline accuracies (see figure 1).

5.5 Selecting Citations

Until now we have considered only the links from one document within the collection to another document within the collection, so $L = J$. Here we investigate the utility of links to documents outside the collections, that is, to documents whose content is unknown; we call these *external links*. Using external links is important as they serve to increase the link density. For reasons of both computational efficiency and information retrieval performance, we use only the most frequently occurring links. This is analogous to the common practice of indexing only the most frequently occurring *terms* in a collection.

It may also be imagined that many of the frequently occurring links on the web would be *stop-links* (like stop-words for terms). These stop-links may be advertisement links or links to search engines which provide no useful information for classification or retrieval. Although stop-links could be removed, robust methods for *stemming* and detecting *stop-links* have yet to be devised. Here we use all the most frequent links.

The documents within the Machine Learning Cora data set cite 34928 documents and in total there are 91842 citations, following an approximately Zipfian distribution [17]. As might be expected, the majority of documents are only cited once (21545 documents). 2970 documents are cited more than 5 times, 1187 documents are cited more than 10 times and only 417 documents are cited more than 20 times. As shown in figure 6, for both collections we used the $L = 200, 500, 1000, 2000$ external and internal documents that are cited most frequently from within the document collection. The details for the Cora and WebKB collections at each L are shown in table 4.

The left hand side of figure 6 shows the marked increase in classification accuracies achieved by using external links for the Cora data. Note also that best classification is obtained when more weight ($0.5 \leq \alpha \leq 0.7$) is given to link information compared with baseline ($\alpha = 0.8$).

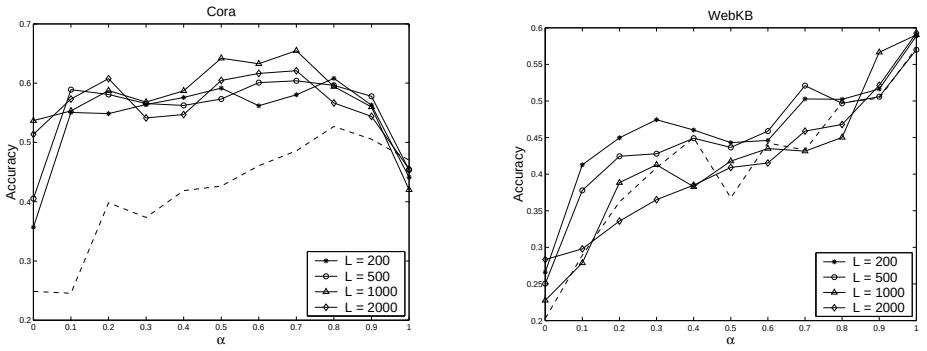

Fig. 6. Average classification accuracy for the Cora (left) and WebKB (right) data sets using the 200, 500, 1000 and 2000 most frequent internal and external links. The dotted lines show the baseline accuracies (see figure 1).

Table 4. Details corresponding to figure 6. *Accuracy* is the mean accuracies over all α and *internal links* is the number of internal links.

	Cora				*WebKB*			
L	$\Gamma \times 10^{-3}$	ρ	*Accuracy*	*Internal links*	$\Gamma \times 10^{-3}$	ρ	*Accuracy*	*Internal links*
200	13.85	2.77	0.540	40/200	6.84	1.37	0.460	43/200
500	8.74	4.38	0.555	113/500	4.06	2.03	0.447	88/500
1000	6.05	6.06	0.575	218/1000	2.68	2.69	0.416	168/1000
2000	4.08	8.18	0.562	418/2000	1.77	3.56	0.412	252/2000

The documents within the WebKB document set cite 53228 documents and in total there are 75958 citations. 44373 of the documents are only cited once from within the collection. 1009 documents are cited in more than 4 documents and only 264 documents are cited more than 10 times.

As shown in figure 6, the classification accuracy for WebKB using the most frequent internal and external links is not improved above the baseline. We believe this is because, by selecting citations we were only able to double the average number of links per page and in doing this we quartered the number of documents that could be cited. The WebKB data set does not contain enough high quality link information for link analysis methods to be useful.

For the Cora collection, as the number of most frequent links is increased (and the link density is decreased) the mean classification accuracy increases until $L = 1000$, but then starts to decrease at $L = 2000$ (this decrease continues with increasing L although not shown in figure 6 or table 4). In contrast, as the number of most frequent links for the WebKB data is increased the mean classification accuracy simply decreases. This effect is due to the difference in link quality in the datasets. The high quality Cora links result in classification accuracy increases as long as the new links provide sufficient information. However, the classification accuracy can be seen to decrease when the noise provided

by the new links outweighs the information. In the case of the WebKB even $L = 200$ links provide more noise than useful information. This effect is also found when selecting terms, using all terms in a collection will decrease performance because the low frequency terms add noise, so the best performance is found when a small number of the most frequently occurring terms is selected.

6 Discussion

We have investigated the utility of links and citations for information access tasks, in particular text classification. The PLSA+PHITS model [3] provides a principled probabilistic model for document-term and document-citation pairs and has been shown here and in [3] to provide improved classification performance when account is taken of the links between documents as well as the terms within them.

The results obtained here indicate better classification accuracy than those reported in [3] on the Cora and WebKB datasets. Note, however, that different numbers of terms and citations were indexed and the TEM training regime differed. Additional improvements can also be obtained by using more latent classes than the number of manually assigned classes, because the document collection pertains to a richer variety of topics than the rather coarse manual classification. Although not shown here, we found that classification is also enhanced by using a k-NN classifier instead of merely the single nearest neighbour. For both the Cora and the WebKB datasets the optimum k, (usually $k_{opt} \approx 15$), achieved an increase in classification of about 10 percent for all α. We anticipate that other more sophisticated classifiers will also be effective, although perhaps at additional computational cost.

Our experiments show that the density of links within a collection is important; indeed, diluting the links in the Cora collection diminishes the classification performance to the point where inclusion of link information always degrades content-based classification. Nonetheless, our experiments with link randomisation and link concentration, show that quality of links is also important. As might be expected, intra-class links are significant for document classification; artificially boosting the intra-class links in the WebKB collection allows link information to enhance, rather than degrade, document classification. We have also shown that utilising links to documents external to a collection (whose content is therefore not known) can improve classification. However, the external links must be of high quality. An important area for future investigation will be the detection of *stop-links* and methods for *stemming* links.

As mentioned in the introduction, TREC web tracks have found little benefit in using content and link information over content methods alone. The TREC-8 Web Track used the WT2g collection which has already been criticised [9] for its low link density and lack of cross-host links. In fact, in 247491 documents there are only 2797 cross-host links. The overall graph sparseness is $\Gamma = 1.9 \times 10^{-5}$; that is, a factor of 34 times lower than the Cora data and 12 times lower than the WebKB data. The results presented here indicate that, at least for document

classification purposes the WT2g link density is far too low to effectively augment content based methods.

Figures for the TREC-9 Web Track collection, which was specifically engineered to be representative of the Web [2], or for the Web as a whole are not available. However, it is unlikely that the density of high quality links approaches that of the Cora collection in which link information is certainly useful. Information access techniques in relatively haphazard collections such as the Web in its current form may therefore have to rely heavily on content.

Acknowledgments

Michelle Fisher is supported by a CASE studentship with BT and the EPSRC. We are grateful for helpful discussions with Gareth Jones.

References

1. CMU world wide knowledge base WebKB project. `www-2.cs.cmu.edu/~webkb/`.
2. P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*, 2001.
3. D. Cohn and T. Hofmann. The missing link – a probabilistic model of document content and hypertext connectivity. *Neural Information Processing Systems*, 13:430–436, 2001. T. Leen et al. eds.
4. N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proc. 24th SIGIR*, pages 250–257, 2001.
5. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. Am. Soc. Info. Science 41*, 6:391–407, 1990.
6. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm with discussion. *Journal Royal Statisical Society 2*, 39:1–38, 1977.
7. Google. `http://www.google.com/technology/whyuse.html`.
8. D. Hawking. Overview of the TREC-9 Web Track. *In 9th Text REtrieval Conference (TREC-9)*, 2000.
9. D. Hawking, E. Voorhees, N. Craswell, and P. Bailey. Overview of the TREC-8 Web Track. *In Eighth Text REtrieval Conference (TREC-8), Gaithersburg, Maryland*, 1999.
10. T. Hofmann. Probabilistic latent semantic indexing. In *Proc. 22nd SIGIR*, pages 50–57, 1999.
11. T. Hofmann and J. Puzicha. Unsupervised learning from dyadic data. Technical Report TR-98-042, University of California, Berkeley, CA, 1998.
12. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
13. A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, 3:127–163, 2000. `http://www.research.whizbang.com/data/`.
14. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.
15. D. Sullivan. Search engine watch, 2002. `www.searchenginewatch.com`.
16. Text REtrieval Conference (TREC) Home Page. `http://www.trec.nist.gov/`.
17. H. Zipf. *Human behaviour and the principle of least effort*. Addison-Wesley, 1949.