Inf Retrieval
DOI 10.1007/s10791-010-9127-7

Discriminative graphical models for faculty homepage discovery

Yi Fang · Luo Si · Aditya P. Mathur

Received: 7 July 2009 / Accepted: 25 January 2010

Abstract Faculty homepage discovery is an important step toward building an academic portal. Although the general homepage finding tasks have been well studied (e.g., TREC-2001 Web Track), faculty homepage discovery has its own special characteristics and not much focused research has been conducted for this task. In this paper, we view faculty homepage discovery as text categorization problems by utilizing Yahoo BOSS API to generate a small list of high-quality candidate homepages. Because the labels of these pages are not independent, standard text categorization methods such as logistic regression, which classify each page separately, are not well suited for this task. By defining homepage dependence graph, we propose a conditional undirected graphical model to make joint predictions by capturing the dependence of the decisions on all the candidate pages. Three cases of dependencies among faculty candidate homepages are considered for constructing the graphical model. Our model utilizes a discriminative approach so that any informative features can be used conveniently. Learning and inference can be done relatively efficiently for the joint prediction model because the homepage dependence graphs resulting from the three cases of dependencies are not densely connected. An extensive set of experiments have been conducted on two testbeds to show the effectiveness of the proposed discriminative graphical model.

Keywords Discriminative graphical models · Homepage finding · Information retrieval

Y. Fang (✉) · L. Si · A. P. Mathur
Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA
e-mail: fangy@cs.purdue.edu

L. Si
e-mail: lsi@cs.purdue.edu

A. P. Mathur
e-mail: apm@cs.purdue.edu

Published online: 19 February 2010

1 Introduction

In response to the complex needs of academic users, many academic information retrieval and management systems have emerged recently. Notable examples include Academia.edu[1], DBLife[2], ArnetMiner[3] and INDURE[4]. These academic portals tend to identify and extract relevant information from the Web without human intervention due to the large scale of information being generated from academic communities. Among many diverse information sources, faculty personal homepages are a valuable and reliable one since they usually contain the up-to-date and relatively well-formatted data about faculty members. Therefore, automatic discovery of faculty homepages is an important step toward building an academic portal.

From the perspective of end users, current commercial search engines work reasonably well for locating potential faculty homepages. When a researcher's name along with his/her institution is searched through Google or Yahoo, it is very likely that the homepage appears among the top returned results if it exists. Exploiting the results from general search engines is a valuable approach to obtain accurate results of faculty homepage discovery.

In this paper, we utilize Yahoo BOSS API to obtain a small set of candidate faculty homepages. Homepage discovery is then reduced to a text classification problem on these high-quality candidate documents. We can then use logistic regression to classify these web pages. However, like most of the current homepage classification methods, logistic regression makes predictions only based on individual web pages, and does not utilize the relations to the other web pages. In fact, these candidate pages are related to each other and their labels are not independent. By defining homepage dependence graph, we propose to use conditional undirected graphical models to capture the dependence of the labels in order to make joint predictions. The dependencies we consider come from three cases: (1) when pages are linked, they are likely to have the same label (e.g., the homepages of advisors and advisees or research collaborators); (2) if one page is a candidate for two or more faculty members, that page is less likely to be the homepages simultaneously for them; (3) if some page is already determined as a true positive, the possibilities of the other candidate pages (for the same people) being homepages should be reduced, because the majority of faculty members have only one homepage. These dependencies can be naturally encoded in the proposed graphical models. Learning and inference can be done relatively efficiently because the resulting homepage dependence graphs are not heavily connected. We follow a discriminative approach, where we optimize the conditional likelihood of the labels given the features. Hence any informative features can be used conveniently. To the best of our knowledge, this is the first research work that applies conditional undirected graphical models to faculty homepage discovery.

In the experiments, we compare heuristics, logistic regression, support vector machine (SVM) and the proposed discriminative graphical models on two data collections with different characteristics. One set of general heuristic rules and two sets of domain specific rules are designed for comparison. The features adopted for discriminative learning are extracted from three sources: page content, URL and links. While the general heuristics are shown more robust across different datasets than the domain-specific heuristics, logistic

[1] http://www.academia.edu/

[2] http://www.dblife.cs.wisc.edu/

[3] http://www.arnetminer.org/

[4] http://www.indure.org/

regression is shown to be more effective than the best heuristics. Furthermore, the proposed joint prediction models outperform logistic regression with a statistically significant difference in both within-institution and cross-institution evaluations. We also show how the proposed model behaves under various amount of training data.

The next section discusses related work. Section 3 describes the proposed joint probabilistic model. Section 4 explains our experimental methodology. Section 5 presents the experimental results and the corresponding discussions. Finally Sect. 6 gives conclusions and discusses future work.

2 Related work

The goal of faculty homepage discovery is to automatically and accurately identify faculty homepages from the Web. Although some related tasks are well studied in the information retrieval community, there are not much focused research devoted to this problem. An early effort for personal homepage discovery can be dated back to 1996 resulting in a popular real-word system called Ahoy (Shakes et al. 1997), which used three external systems and applied simple yet successful heuristics to retrieve the users homepage. Ahoy was presented in 1997, and many of components that Ahoy was performing are now embedded into commercial search engines. Another similar homepage finding system carried out at HP was designed to search for personal homepages of computer scientists. In April 2000 more than 42,000 entries were stored there, but the service has ceased to operate.

Homepage finding or entry page search is a closely related task but with a more general scope. It has attracted increasing attention in the IR research community since its first run as part of TREC-10 Web track (Hawking and Craswell 2001; Voorhees and Harman 2001). The goal was to find the entry page of the site described in the topic. Kraaij et al. (2002) obtained among the best results for the TREC-10 homepage finding task by assigning a prior probability for a page being a homepage. Craswell et al. (2001) found anchor text retrieval is more effective than full-text retrieval for the homepage finding task. They also adopt mean reciprocal rank as the evaluation measure, which has become widely used for homepage finding task. Upstill et al. (2003) has shown that the query-independent evidence is beneficial on homepage finding. Experiments of other groups confirmed the effectiveness of these features (Xi et al. 2002; Craswell et al. 2002). Ogilvie and Callan (2003) have tried to make use of different document representations from different sources in language models for homepage finding tasks.

Davison (2000) empirically demonstrated that the topic locality assumption holds: most web pages are linked to others with related content. Although link structure analysis has not been able to improve retrieval effectiveness for ad hoc search tasks, the TREC-2001 evaluation (Westerveld et al. 2002) showed that it does improve performance in a homepage finding task. The idea is to utilize the fact that entry pages tend to have a higher number of incoming links than other documents. However, the locality of topics has not yet been fully exploited by making joint discovery through the dependence of linked homepages.

Academic portals view faculty homepage discovery as an essential component to collect faculty information. Some of the previous research tried to directly apply text categorization techniques (Sebastiani 2002; Yang and Pedersen 1997) to the task. For example, in DBLife a set of candidate web pages are collected, and then SVM is trained and applied to classify these pages (Doan et al. 2006). ArnetMiner also employs the same procedure for its faculty homepage discovery task (Tang et al. 2007). Culotta et al. (2004) primarily applies a set of heuristic techniques to filter the results from Google. However, these

approaches ignore the relations among the labels of candidate homepages and they classify each page independently.

In the data mining community, some collective classification models (Neville and Jensen 2003; Heckerman et al. 2007; Taskar et al. 2002) have been developed to exploit the dependencies in related entities to improve predictions. These models generally follow the framework of undirected graphical models (Jordan 1998) or conditional random fields (Lafferty et al. 2001). By modeling relational dependencies, they have been shown to achieve significant improvements over independent predictions. However, none of them is specifically designed for faculty homepage discovery by considering special characteristics of the problem.

3 Faculty homepage discovery

3.1 Yahoo BOSS API

BOSS[5] (Build Your Own Service) is Yahoo's open search web services platform that enables developers to easily find and manipulate information on the Web. We can issue search requests along with a set of parameters to Yahoo's index of billions of web pages and receive results as structured data. BOSS offers developers unlimited daily queries and supports the same search syntax as the Yahoo.com site. BOSS has built-in functions which can extract title, URL and snippet from a web page returned as part of a search request. In this paper, the query we used for each faculty member consists of his/her full name and institution (e.g., "Elisa Bertino Purdue"), along with a set of filtering rules such as excluding non-HTML file and only focusing on the .edu domain(the heuristic can be easily adjusted for academics outside the USA). We collect the top returned results as candidate homepages. Because of the high likelihood of these pages being homepages, the classification based on them is expected to yield high accuracy.

3.2 Individual prediction by logistic regression

Logistic regression is a widely used classification method, which makes predictions separately for individual entities. In the context of homepage categorization, a text classifier is learned from a set of training pages $D = \{(x_1; t_1),\ldots,(x_n; t_n)\}$ where x_i denotes the feature vector of the ith page and $t_i \in \{0, 1\}$ is the corresponding label. Because the candidate homepages are retrieved by corresponding faculty names using the search engine, every retrieved page is associated with a faculty member Y. As a result, the classifier is to decide whether the page is Y's homepage. We use the parametric form of logistic regression to model the probability of relevance of a page as $P(t|x)$:

$$P(t = 1|x) = \frac{1}{1 + \exp(\theta' f(x))}$$

$$P(t = 0|x) = \frac{\exp(\theta' f(x))}{1 + \exp(\theta' f(x))}$$

[5] http://www.developer.yahoo.com/search/boss. The version of API we used is called BOSS-U, the academic track of the BOSS initiative.

where $f(x)$ is a feature vector of the web page x, and θ is the corresponding weight vector. Training (i.e., finding the parameter θ) can be done by maximizing the conditional log likelihood of training data

$$\sum_{i=1}^{n} \log P(t_i|x_i;\theta)$$

This is a convex function so there is a unique global maximum. While there is no closed form solution, one typically solves the optimization problem with Newton–Raphson iterations, also known as iterative reweighted least squares for logistic regression.

3.3 Homepage label dependence

Logistic regression model is one of the most effective techniques for general text categorization problems (Yang and Liu 1999). However, it makes decision only based on the features of individual web pages and ignores the dependencies between their labels. In fact, the labels of web pages are not independent and are correlated with each other. In this paper, we consider three cases of dependencies (called Case A, Case B and Case C dependence respectively). First of all, by the topical locality effect, when pages are linked, they are likely to contain related contents. It means that web pages that are linked to faculty homepages are potentially relevant homepages themselves (e.g., the linked homepages of advisors and advisees or research collaborators). Secondly, if one page is a candidate for two or more faculty members, that page is less likely to be the homepages simultaneously for them, because normally people have their own homepages. On the other hand, we cannot completely eliminate the possibility that a single homepage cannot be shared among multiple persons because in some cases people do list group page as their homepage. Thirdly, because faculty are less likely to have multiple homepages, if some page is already determined as a true positive, the possibilities of the other candidate pages (for the same people) being homepages should be reduced. Therefore, the top n pages returned by the same query are mutually dependent. In all the three cases, the dependencies have uncertainty involved and it is more suitable to use probability than rigid rules to represent and manipulate the dependence.

Figure 1 illustrates the examples of the three cases. In Fig. 1a, Prof. Michael Jordan's homepage links to Prof. David Blei's. Therefore, if one of them is known as a true homepage, another one becomes more likely to be a homepage. Figure 1a′ ignores the directions of the edges in Fig. 1a. This representation is useful for our undirected graphical models defined in

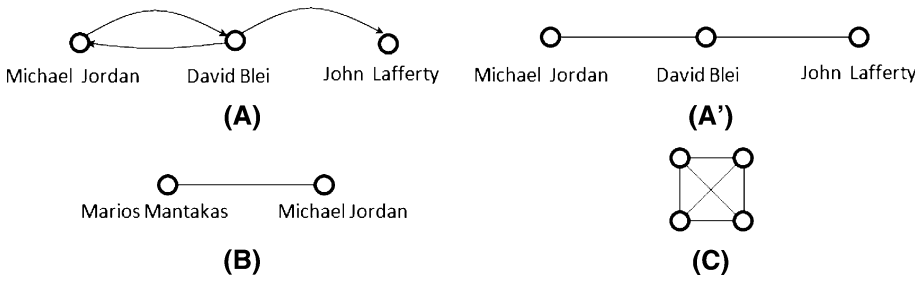

Fig. 1 Three cases of dependence of labeling faculty homepages: **a** Case A dependence with directional edges representing hypertext links from source to destination; **a′** Case A dependence ignoring the directions of edges; **b** Case B dependence; **c** Case C dependence with four candidate homepages for each faculty member

the next section. In Fig. 1b, Yahoo returns Prof. Michael Jordan's homepage as the top result in response to the query for Prof. Marios Mantakas. If we use independent prediction, this page is very likely to be classified as homepage for both professors (because it really is a homepage). But if the page is already known as Prof. Jordan's homepage, the probability of the page being another professor's homepage should be reduced. Figure 1c represents the top four results returned from Yahoo for a query. As we discussed in Case C dependence, once one of them is identified as a homepage, the others should be less likely to be homepages. Therefore, these 4 web pages are mutually dependent with edges connecting to each other. This paper focuses on these three types of dependent relationships, but the work can be naturally expanded to incorporate other types of dependence.

3.4 Joint prediction by discriminate graphical models

It is quite intuitive that the dependencies among labels of candidate homepages could help us make predictions. Then how to capture the dependencies and encode them into the prediction models becomes a central question. We give the following formal representation of the dependencies in graphical models. An undirected graph $G = <V, E>$, called homepage dependence graph (HDG), can be defined for the set of selected candidate homepages, where $V = T \cup X$ and $E = E_T \cup E_X$. Specifically, $t_i \in T$ is the node representing the label of the web page i and $x_i \in X$ is the node to represent observed attributes for the page. E_X is the set of edges showing the relations between observed page attributes and their labels. $E_T = A \cup B \cup C$ is the set of edges indicating dependencies between web page labels. Corresponding to the three cases illustrated in Fig. 1, we denote $e_{kj} \in A$ if there exists a link between t_k and t_j regardless of directions, $e_{kj} \in B$ if t_k and t_j are the same page and $e_{kj} \in C$ if both t_k and t_j are returned for the same query. Figure 2 shows an instance of the defined model with three faculty members having four candidate homepages for each.

A clique, which is an important concept in HDG, is defined as a set of vertices S that $\forall u$, $v \in S$, there exists an edge connecting u and v. In other words, the subgraph induced by a clique is a complete graph. In Case C dependence, the top n pages returned for the same query/faculty member are mutually dependent and thus form a clique, which is denoted by CL_r for the faculty member r. In a trivial example, $\{t_6, t_{11}\}$ in Fig. 2 also forms a clique of two vertices. In defining an undirected graphical model, one can define arbitrary potential functions on cliques of arbitrary sizes. However, large cliques are problematic both for computational and statistical reasons, because the inference complexity is exponential to the maximum clique size, and estimation of large number of parameters requires a large

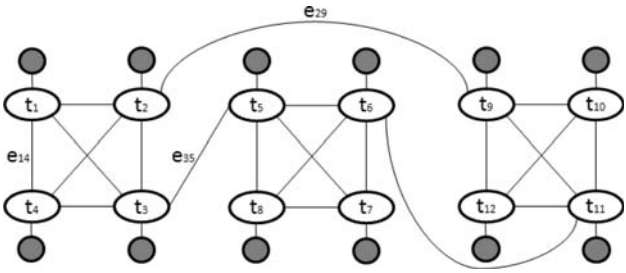

Fig. 2 An instance of homepage dependence graph with three faculty members having four candidate homepages returned for each. The nodes X of observed page attributes are shaded. This example includes three cases of dependence (e.g., $e_{29} \in A$, $e_{35} \in B$ and $e_{14} \in C$)

amount of training data. Fortunately, the three cases of dependence we considered between the labels result in relatively sparse HDGs in real applications.

Given the homepage dependence graph G defined above, following the idea of conditional random fields (Lafferty et al. 2001), the conditional probability of label set T given the observation X is defined as follows:

$$P(T|X) = \frac{1}{Z}\exp\left(\sum_i \psi_1(x_i, t_i) + \sum_{e_{kj}\in A} \psi_2(t_k, t_j, x_k, x_j) + \sum_{e_{kj}\in B} \psi_3(t_k, t_j, x_k, x_j) + \sum_r \psi_4(CL_r)\right)$$

where the exponential function ensures that $P(T|X)$ is positive, and the normalizing constant defined below

$$Z = \sum_{t'}\left(\exp\left(\sum_i \psi_1(x_i, t'_i) + \sum_{e_{kj}\in A} \psi_2(t'_k, t'_j, x_k, x_j) + \sum_{e_{kj}\in B} \psi_3(t'_k, t'_j, x_k, x_j) + \sum_r \psi_4(CL_r)\right)\right)$$

ensures that $P(T|X)$ sums to 1. $\exp(\psi_i)$ represents potential functions in which ψ_1 captures the degree to which t_i is compatible with x_i. ψ_2 and ψ_3 captures the degree to which t_k is compatible with t_j for Case A and Case B dependence defined in Sect. 3.3 respectively, and ψ_4 is defined on the clique CL_r for Case C dependence. ψ_i can be arbitrary real-valued functions. In particular, considering that pages with the same label tend to link to each other, we can capture this correlation by defining ψ_2 on the clique having higher values for assignments that give a common label to the linked pages. For ψ_3, the edge $e_{kj} \in B$ suggests that a conflicting label assignment for t_k and t_j should have higher values. Similarly, ψ_4 should receive high value when there is only one positive label in the clique and otherwise get low values especially when all the labels are positive. Generally, the functions $\{\psi_1, \psi_2, \psi_3, \psi_4\}$ are often represented by weighted combinations of feature functions in the following form:

$$\psi_1(x_i, t_i) = \sum_{m=1}^{M_1} \alpha_m f_m(x_i, t_i)$$

$$\psi_2(t_k, t_j, x_k, x_j) = \sum_{m=1}^{M_2} \beta_m g_m(t_k, t_j, x_k, x_j)$$

$$\psi_3(t_k, t_j, x_k, x_j) = \sum_{m=1}^{M_3} \gamma_m h_m(t_k, t_j, x_k, x_j)$$

$$\psi_4(CL_r) = \sum_{m=1}^{M_4} \eta_m s_m(CL_r)$$

where α_m, β_m, γ_m and η_m are trainable weights, f_m, g_m, h_m and s_m are the feature functions and M_1, M_2, M_3 and M_4 are the corresponding numbers of features. This formulation of the discriminative undirected graphical models is completely general and flexible into which any potential functions or features can be encoded. Once a parameterization is chosen, the graphical model can be trained to maximize the log

likelihood of the training data as max$\log P(T|X)$. To help prevent overfitting, we maximize a posteriori estimation by defining priors such as $P(\alpha_m) = \frac{1}{2\pi\sigma_\alpha}\exp(-\frac{\alpha_m^2}{\sigma_\alpha^2})$ where σ_α is a hyperparameter which can be determined empirically or by cross-validation. As a result, the following regularized objective function is maximized:

$$\begin{aligned}
J(\alpha,\beta,\gamma,\eta) &= \log P(T|X) - \sum_{m=1}^{M_1}\frac{\alpha_m^2}{2\sigma_\alpha^2} - \sum_{m=1}^{M_2}\frac{\beta_m^2}{2\sigma_\beta^2} - \sum_{m=1}^{M_3}\frac{\gamma_m^2}{2\sigma_\gamma^2} - \sum_{m=1}^{M_4}\frac{\eta_m^2}{2\sigma_\eta^2} \\
&= \log\frac{1}{\mathcal{Z}}\exp\left(\sum_{i=1}^{n}\psi_1(x_i,t_i) + \sum_{e_{kj}\in A}\psi_2(t_k,t_j,x_k,x_j) + \sum_{e_{kj}\in B}\psi_3(t_k,t_j,x_k,x_j)\right. \\
&\qquad \left. + \sum_r \psi_4(CL_r)\right) - \sum_{m=1}^{M_1}\frac{\alpha_m^2}{2\sigma_\alpha^2} - \sum_{m=1}^{M_2}\frac{\beta_m^2}{2\sigma_\beta^2} - \sum_{m=1}^{M_3}\frac{\gamma_m^2}{2\sigma_\gamma^2} - \sum_{m=1}^{M_4}\frac{\eta_m^2}{2\sigma_\eta^2} \\
&= \sum_{i=1}^{n}\psi_1(x_i,t_i) + \sum_{e_{kj}\in A}\psi_2(t_k,t_j,x_k,x_j) + \sum_{e_{kj}\in B}\psi_3(t_k,t_j,x_k,x_j) \\
&\quad + \sum_r \psi_4(CL_r) - \log Z - \sum_{m=1}^{M_1}\frac{\alpha_m^2}{2\sigma_\alpha^2} - \sum_{m=1}^{M_2}\frac{\beta_m^2}{2\sigma_\beta^2} - \sum_{m=1}^{M_3}\frac{\gamma_m^2}{2\sigma_\gamma^2} - \sum_{m=1}^{M_4}\frac{\eta_m^2}{2\sigma_\eta^2}
\end{aligned}$$

There is no closed form solution to the maximization problem above, therefore we use the iterative searching algorithm BFGS (McCallum 2003). It is worth noticing that logistic regression is actually a special case of the above graphical model where there is no edges among labels T. By taking the first derivative of the log likelihood $J(\alpha, \beta, \gamma, \eta)$ with respect to the parameter α_m, we have

$$\frac{\partial J}{\partial \alpha_m} = \sum_{i=1}^{n}\left(f_m(x_i,t_i) - E_{P(T)}[f_m(x_i,t_i)]\right) - \frac{\alpha_m}{\sigma_\alpha^2}$$

The first two terms are the difference between the expected feature counts and the empirical feature counts, where the expectation is taken relative to the marginal distribution $P(T)$:

$$E_{P(T)}[f_m(x_i,t_i)] = \sum_{t_i'} f_m(x_i,t_i')P(t_i'|X)$$

Similarly, for β_m, γ_m and η_m, we have

$$\begin{aligned}
\frac{\partial J}{\partial \beta_m} &= \sum_{e_{kj}\in A}\left(g_m(t_k,t_j,x_k,x_j) - E_{P(T)}[g_m(t_k,t_j,x_k,x_j)]\right) - \frac{\beta_m}{\sigma_\beta^2} \\
\frac{\partial J}{\partial \gamma_m} &= \sum_{e_{kj}\in B}\left(h_m(t_k,t_j,x_k,x_j) - E_{P(T)}[h_m(t_k,t_j,x_k,x_j)]\right) - \frac{\gamma_m}{\sigma_\gamma^2} \\
\frac{\partial J}{\partial \eta_m} &= \sum_{r}\left(s_m(CL_r) - E_{P(T)}[s_m(CL_r)]\right) - \frac{\eta_m}{\sigma_\eta^2}
\end{aligned}$$

The prediction task in our conditional graphical model is to compute the posterior distribution over the label variables T given the feature variables X, i.e., to compute the following most probable assignment:

$$T^* = \arg\max_T P(T|X)$$

Exact algorithms for inference in graphical models can be done efficiently for specific graph topologies such as sequences and trees. However, our homepage dependence graphs go beyond these simple topologies and exact inference is usually intractable in this case. Therefore, we resort to belief propagation (BP) (Pearl 1988) for its simplicity, relative efficiency and accuracy. BP is a message passing algorithm for performing inference on graphical models. It does exact marginalization in tree structured graphs and can also be applied to graphs with loops to get approximate marginals (Yedidia et al. 2001; Murphy et al. 1999). Fortunately, the homepage dependence graphs resulting from the three cases of dependence are usually not densely connected in real applications. Therefore, the inference task can be done relatively efficiently by BP for the proposed graphical models.

The conditional undirected graphical models defined above are well applicable to our faculty homepage discovery task. First, in contrast to directed models, the undirected models do not impose the acyclicity constraint that hinders the representation of Case B and Case C dependencies. Second, the conditional models are well suited for discriminative training, where we optimize the conditional likelihood of the labels given the features, which generally improves classification accuracy over their generative counterparts.

4 Experiments

4.1 Data

We test our models on two data collections from Purdue university and Indiana university (IU) respectively. 1,000 faculty members are randomly chosen from each university. For each faculty member, we select the top 4 results (if there exist) returned from Yahoo. In INDURE, some faculty homepages are available for each institution as they are uploaded into the system by faculty themselves or their department heads. These pages can be utilized as the labeled pages. The rest pages from Yahoo are manually annotated. Some faculty have pages on the department website that list their detailed information such as eduction, publications and funded projects. These pages are also deemed as homepages. Thus, faculty may have more than one homepage. In one experiment of Sect. 5.3, we also report the results on the faculty personal homepages only which exclude the obligatory department homepages.

Table 1 gives detailed statistics of the data collections. We can notice that the two datasets have different characteristics. In particular, the number of faculty who do not have any homepage is larger in IU than in Purdue. It may come from the fact that Purdue has larger science and engineering programs in which faculty members are more willing to create their homepages. This may also explain some other numbers in the table such as the number of faculty who have more than 1 homepage and the numbers of edges in Case A and Case B. Because the total number of selected faculty is relatively small and they are restricted to single institutions, we can find there are not too many links between them, but we will see in Sect. 6 that these small number of links could make a difference. On the other hand, we can see the maximum clique size in both universities is four, which indicates that the homepage dependence graphs are not densely connected.

In the experiments, we evaluated the proposed model under two settings: within-institution testing and cross-institution testing. In within-institution testing, we split the data of one institution to training and testing. Specifically, for each dataset, we carried out two-fold cross-validations by splitting 1,000 faculty members into 500 for training and 500 for testing in each fold. Therefore, testing is done on the same institution with training and

Table 1 Statistics of test collections

Collection	Total # of faculty	Total # of web pages	Average words/page	SD of page size
Purdue	1000	3981	312.7	688.9
Indiana	1000	3957	376.5	794.3
	# of faculty having more than 1 homepage	# of faculty having no homepage	# of edges in Case A	Average degree of vertices
Purdue	132	177	169	3.11
Indiana	71	295	106	3.09
	# of faculty having <4 candidate homepages	Maximum degree of vertices	# of edges in Case B	Maximum clique size
Purdue	17	11	32	4
Indiana	38	7	48	4

thus the learned homepage classifier is specific for each institution. Within-institution testing is not applicable in some scenarios where there are no training data available for specific institutions. In contrast, in cross-institution testing, we use the data of one institution for training and test the learned model on another institution. Sections 5.1, 5.2 and 5.4 includes the results by within-institution evaluation, and Sect. 5.3 includes the results by cross-institution evaluation. In both settings, across-network classification is applied: learning from one network and applying the learned model to another network. To evaluate the methods, precision(p) and recall(r) were adopted in their combined form $F1$, which is defined as $F1 = \frac{2rp}{r+p}$. Although mean reciprocal rank or MRR is widely used as an evaluation measure for general homepage finding tasks, it may not be suitable for the purpose of building an academic portal because MRR is usually used to evaluate a ranked list of results instead of a set of classified items.

4.2 Experiment setup

In the experiments, we compared heuristic rules, logistic regression, SVM and the joint prediction approaches. The heuristic rules method is similar to decision tree models but with manually tuned decision rules. The heuristics we used are based on a number of observations that are often yet not always true for faculty homepages. Figure 3 shows a part of our heuristic rules.

While these are general heuristic rules, there also exist university-specific rules. For example, many top returned results from Yahoo for Purdue faculty come from Purdue News Service news[6] where significant research discoveries from faculty are often reported. Therefore, we can filter out all the web pages from this news service website. Another good heuristic for Purdue is that the URL length of its faculty homepages is usually short (less than 60), but this rule does not hold for IU people. Therefore, the sets of university-specific

[6] http://www.news.uns.purdue.edu/

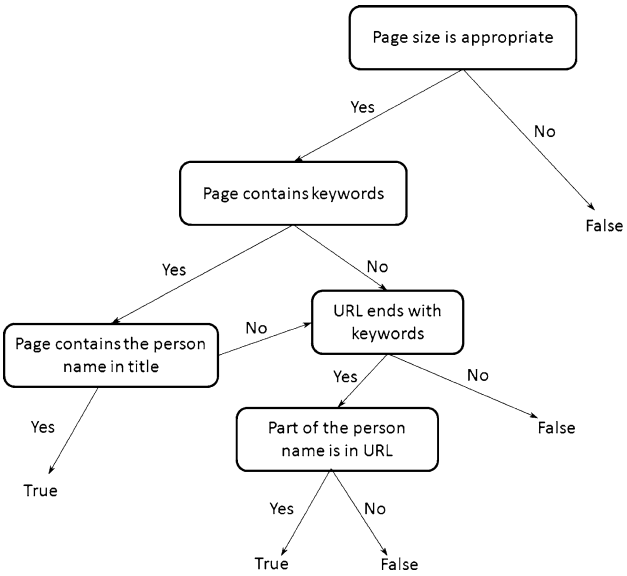

Fig. 3 An example set of heuristic rules

Table 2 Part of page attribute features organized in three types used by logistic regression, SVM and the joint prediction model

Content	Page size
	Anchor text
	Title field
	Whether the page contains a set of predefined keywords
	Rank in Google
URL	Number of slashes in the URL
	Whether the URL contains a set of keywords
	String distance between faculty name and the last part of URL
Links	Number of outgoing links of the web page
	Number of outgoing links to outer domain
	Number of outgoing links to the same domain
	The above scores normalized by page size

heuristics may be conflicting with each other. In the experiments, we designed three sets of heuristic rules for comparison, which are general rules (Rul_1), Purdue specific rules (Rul_2) and IU specific rules (Rul_3), respectively.

For logistic regression and conditional graphical models, we select a set of page attribute features (i.e., X) for the discriminative training. As presented in Table 2, the total features are divided into three groups based on where the features come from: content, URL and link. The content keywords include some indicative words such as "homepage", "research", "education", "publication" and the particular faculty name as well. The URL keywords include "home", "index", "faculty" and so on. There are totally 52 features used in our experiments and all the feature scores are normalized by the maximum score in

that feature. Since the focus of this study is on the probabilistic models rather than feature engineering, we do not intend to choose a comprehensive set of features. It is noted that the features g_m, h_m and s_m (usually binary) defined in Sect. 3.4 have dependence on the page attribute features X, but in the experiments we simplify these feature functions by dropping the dependence (while f_m still depends on X). For example, the feature function g_1 is defined as follows: $g_1 = 1$ if $t_k = 1$ and $t_j = 1$; otherwise $g_1 = 0$. As a result, these features are only dependent on the labels of related pages and thus the number of parameters to be estimated is reduced.

5 Results

An extensive set of experiments were conducted on the two testbeds to address the following questions:

(1) How good are machine learning approaches, in particular, logistic regression, compared with heuristics? The experiments described in Sect. 5.1 are conducted on three sets of heuristic rules as well as logistic regression with different combinations of features.
(2) Can the prediction performance be improved by considering the dependence between the labels of candidate homepages? Experiments in Sect. 5.2 are conducted to compare the proposed joint probabilistic model with logistic regression and SVM. The joint prediction model with different cases of dependence are also compared.
(3) How does the proposed joint prediction model perform in the cross-institution evaluation setting? How well does it perform only on the personal faculty homepages? Section 5.3 investigates these questions.
(4) How does the joint prediction model behave with different amount of training data returned from Yahoo? The experiments in Sect. 5.4 are performed to examine how the precision, recall and F1 measures change when various amount of training data are available.

5.1 Heuristic rules versus logistic regression

In this experiment, we compare the heuristic methods described in Sect. 4.2 with the logistic regression (LR) model. Table 3 contains the comparisons in F1 score. We also report the F1 scores obtained by the approach (denoted by *SE*) that always takes the top result from the search engine as the homepage. We can see that the heuristic rules can improve upon the *SE* approach, and so can logistic regression especially when the content

Table 3 Comparison of F1-score between heuristic rule based approach and logistic regression on Purdue and Indiana datasets.

Collection	SE	Rul_1	Rul_2	Rul_3	LR_1	LR_2	LR_3	LR_{12}	LR_{13}	LR_{23}	LR_{123}
Purdue	0.697	0.759	0.791	0.726	0.796	0.713	0.572	0.844	0.814	0.732	**0.856**
Indiana	0.684	0.753	0.713	0.774	0.785	0.717	0.594	0.832	0.806	0.743	**0.844**

SE denotes the approach that always takes the top result from the search engine as the homepage. Rul_1, Rul_2 and Rul_3 denotes the three heuristic rule-based approaches. *LR* denotes the logistic regression model with the subscript representing different types of features being included (1: content, 2: URL and 3: link). Best results on each collection are highlighted with bold

features are used. We can also find that generally both the heuristic rules and logistic regression approaches perform better on Purdue than on IU. The results for the heuristic approaches are consistent with our expectation that the university specific heuristic rules perform the best among the three on their respective domain data (i.e., Rul_2 for Purdue and Rul_3 for Indiana), while the general rules (i.e., Rul_1) achieve a compromise level of performance between the two sets of domain specific rules. Moreover, logistic regression is tested on different combinations of features. It is not surprising to see that the utilization of all the features yielded the best results, which are also found statistically significant over the heuristic approaches by the sign s-test with α-level of 0.1. The link features (LR_3) alone are weak while the content based features are more indicative (LR_1). In addition, the URL features (LR_2) are quite discriminative given that the total number of them is relatively small.

5.2 Individual classification versus joint prediction

Table 4 contains the within-institution evaluation in F1 score between the individual prediction by LR, SVM[7] and the joint prediction approach. All the three sets of features described in Table 2 are included in this experiment. The table shows that SVM yields comparable results with LR and it is hard to tell which one is better in a general sense, but we can see that both of them are surpassed by the joint prediction models in performance. The graphical models for Case A, Case B, Case C and the combined case are examined respectively. From Table 4 we can see that on both collections, our joint prediction approach consistently outperforms the logistic regression model and SVM in all the scenarios. The model with Case A edges brings more improvement than with Case B, which may be explained by the fact that fewer Case B dependence exists in the data. On the other hand, with more Case A edges, Purdue gains bigger improvement from Case A than IU does. When all the cases of dependencies are considered, the improvements of joint prediction over LR on the two collections are found to be statistically significant by the sign *s*-test with α-level of 0.1.

5.3 Cross-institution evaluation

In this section, the same set of experiments as those in Sect. 5.2 are done but with cross-institution evaluation instead. Table 5 includes the results. In the table, the row of Purdue contains the F1 scores obtained by using the Indiana data as training and then testing on Purdue. Similarly, the row of Indiana is trained on Purdue and tested on Indiana. By comparing LR, SVM and the joint model in the table, we can see that the results generally follow the same pattern with Table 4. On the other hand, if we compare Table 5 with Table 4, we can find that the performance achieved by within-institution testing is better than that by cross-institution. This observation is consistent with our expectation because the training and testing data within the same university usually share more characteristics than those across different universities. Therefore, in within-institution training, the weights of the classifiers are specifically well tuned for specific institutions and thus yield better performance. On the other hand, the relative improvements brought by the joint model in Table 5 are generally more visible than those in Table 4, especially when Case A dependence is considered. This may be explained by the following two reasons: (1) there are less Case A and Case B dependencies in within-institution testing than in cross-institution because cross-institution uses the whole institution data for testing while within-

[7] The SVM^{light} toolkit (http://www.svmlight.joachims.org/) is applied.

Table 4 Comparison of F1-score between logistic regression, SVM and joint probabilistic models by within-institution testing on Purdue and Indiana datasets.

Collection	LR	SVM	%±LR	J_A	%±LR	J_B	%±LR	J_C	%±LR	J_{ABC}	%±LR
Purdue	85.6	85.2	−0.47	88.1	+2.92	86.7	+1.29	87.2	+1.87	**89.1**†	+4.09
Indiana	84.4	84.9	+0.59	86.3	+2.25	85.6	+1.42	86.1	+2.01	**87.5**†	+3.67

LR denotes the logistic regression model, *SVM* denotes the support vector machine classifier, *J* denotes the joint prediction model with subscript denoting different cases of dependence being considered. %±LR denotes relative percentage change over LR approach. Best results on each collection are highlighted with bold. The † symbol indicates statistical significance at 0.9 confidence interval

Table 5 Comparison of F1-score between logistic regression, SVM and joint probabilistic models by cross-institution testing on Purdue and Indiana datasets

Collection	LR	SVM	%±LR	J_A	%±LR	J_B	%±LR	J_C	%±LR	J_{ABC}	%±LR
Purdue	81.7	82.4	+0.86	84.6	+3.55	82.8	+1.35	83.3	+1.96	**85.7**†	+4.90
Indiana	82.1	81.8	−0.37	84.4	+2.80	83.3	+1.46	83.9	+2.19	**85.6**†	+4.26

LR denotes the logistic regression model, *SVM* denotes the support vector machine classifier, *J* denotes the joint prediction model with subscript denoting different cases of dependence being considered. %±LR denotes relative percentage change over *LR* approach. Best results on each collection are highlighted with bold. The † symbol indicates statistical significance at 0.9 confidence interval

Table 6 Comparison of F1-score between logistic regression, SVM and joint probabilistic models by cross-institution testing on the personal faculty homepages only

Collection	LR	SVM	%±LR	J_A	%±LR	J_B	%±LR	J_C	%±LR	J_{ABC}	%±LR
Purdue	82.5	82.9	+0.48	85.8	+4.00	83.7	+1.45	84.1	+1.94	**87.3**†	+5.82
Indiana	82.8	83.1	+0.36	85.3	+3.02	84.1	+1.57	84.7	+2.29	**87.1**†	+5.19

LR denotes the logistic regression model, *SVM* denotes the support vector machine classifier, *J* denotes the joint prediction model with subscript denoting different cases of dependence being considered. %±LR denotes relative percentage change over LR approach. Best results on each collection are highlighted with bold. The † symbol indicates statistical significance at 0.9 confidence interval

institution only uses half of them; (2) in cross-institution testing, page attribute features are less effective to predict labels, and therefore the dependencies between page labels become more prominent features as they are ubiquitous across institutions. These observations suggest the proposed joint prediction model is even more suitable for the cross-institution scenarios where the page attribute features may have large variations across different institutions.

We also report the results on the personal faculty homepages only which exclude the obligatory department homepages. Table 6 contains the results, which generally follow the same pattern as seen in Tables 5 and 4. By comparing Table 6 with Table 5, we can find that the results on the personal homepages are better than on the whole. This may be explained by the fact that the features of personal homepages are stable across institutions while the departmental homepages may depend on specific universities. In addition, we can see that more relative improvements are gained by the joint model on the personal homepages, probably because there are more Case A and Case B dependencies among them than among department homepages.

5.4 Joint prediction models with various amount of training data

To illustrate how our proposed joint prediction model behaves under different amount of training data, we do the experiments in within-institution evaluation on the data from top 1 to top 4 results, respectively, returned from Yahoo. The joint prediction models include all the defined features and all the edges from the three cases. We still use two-fold cross validation to train and test the model for the four settings. Figure 4 shows the resulting F1 scores on the two data collections. Figures 5 and 6 contains the corresponding precision-recall scores. We can see that as more and more data are incorporated, the F1 score generally increases but with less acceleration. After examining Figs. 5 and 6, we may be able to find out the reason. Generally, the recall increases from top 1 to top 4 but with decreasing acceleration of increase, because the top search engine results are much more likely to hit the target than those below them. For both datasets, the increase in F1 score from top 1 to top 2 is mainly contributed by the large increase in recall. From top 2 to top 3,

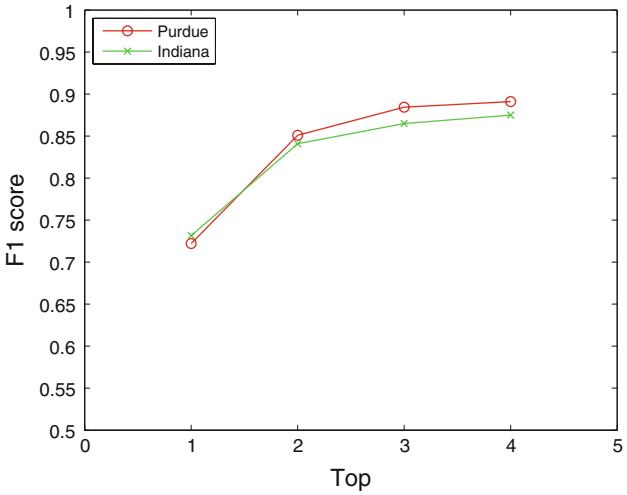

Fig. 4 F1 scores of the joint prediction model on the Purdue and IU faculty members with different amount of data returned from Yahoo

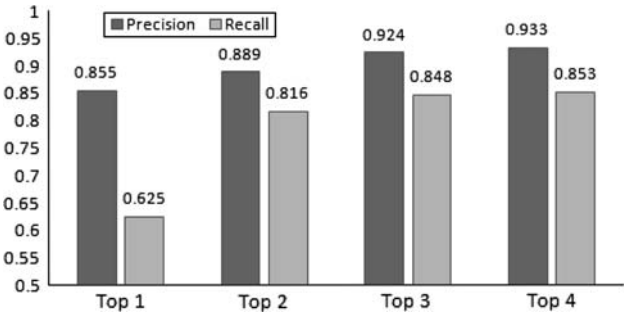

Fig. 5 Precision and recall for the Purdue faculty members with top 1, top 2, top 3 and top 4 returned results, respectively

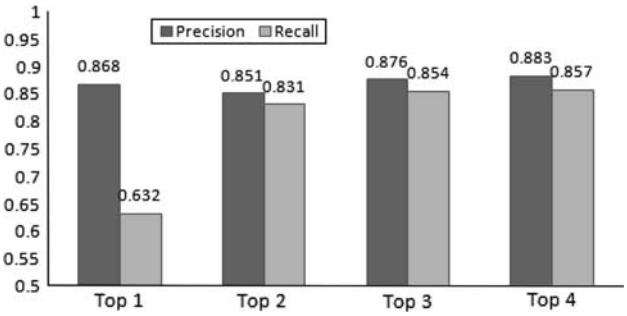

Fig. 6 Precision and recall for the Indiana faculty members with top 1, top 2, top 3 and top 4 returned results, respectively

the precisions of both cases go up probably because more training data are available and more links can be formed. The flat trends in F1 score from top 3 to top 4 suggest that our approach of only selecting data from the top 4 results is sufficient.

An interesting observation between Purdue and IU is that the F1 scores of Purdue are larger than those of IU except the case of top 1. Figure 6 tells us that IU has a higher recall than Purdue at top 1. This may be explained by the fact that fewer IU faculty have homepages and thus the top 1 results returned from the search engine would likely to cover relatively larger percentage of true positive homepages. The coverage gap is narrowed when more candidate pages are obtained.

6 Conclusions and future work

Faculty homepage discovery is crucial for automatic and accurate extraction of faculty information from the Web. It becomes an important domain-specific information retrieval task as more and more academic search engines and portals have been being built. The task is also closely related to other IR problems such as named page finding and topic distillation as presented in TREC Web Track. In this paper, we reduce the task to text categorization problems by utilizing Yahoo BOSS API to select a set of candidate faculty homepages. Previous homepage classification methods make predictions on each pages separately. In contrast, we propose a joint prediction model to simultaneously decide the labels of all the pages together by considering their mutual correlations. The label dependencies can be naturally represented in our homepage dependence graph. Our approach is based on discriminatively trained undirected graphical models which generally improve accuracy over their generative counterparts. The experimental results are provided on two testbeds with different characteristics, showing significant prediction improvements arising from the modeling of relational dependencies.

There are several possibilities to extend the work in this paper. First of all, the proposed approach and three cases of dependencies in this paper are not limited to faculty homepage discovery, but also applicable to homepage finding or entity finding tasks in general such as the problem in the recently launched TREC Entity Track. However, for the general task of entity homepage finding, Case B and Case C dependence is always present while in many scenarios Case A dependence is rarely visible or even does not exist. Instead, we can utilize co-citation rather than citation to describe the dependence (e.g., whether both pages link to an authoritative hub). On the other hand, we can strengthen the proposed discriminative

graphical model by introducing latent variables between observed page features and their labels. These latent variables may be particularly useful when the faculty homepages express strong heterogeneousness (e.g., faculty homepages may come from a diverse ranges of disciplines). The latent variables can reflect the degree to which a web page belonging to the categories and the associated parameters can then be more flexible and better tuned. Furthermore, in many IR applications, we usually need to achieve a balance between precision and recall. In the academic portal applications, a certain degree of leaning toward precision is desirable, because faculty would not like to see their homepage URLs are incorrectly shown. In the future work, we will elaborate the proposed model to improve the precision measure while at the same time not deteriorate recall, in order to maximize user satisfaction for academic portals.

References

Craswell, N., Hawking, D., & Robertson, S. (2001). Effective site finding using link anchor information.

Craswell, N., Hawking, D., Wilkinson, R., & Wu, M. (2002). TREC 10 Web and interactive tracks at CSIRO. NIST special publication, pp. 151–158.

Culotta, A., Bekkerman, R., & McCallum, A. (2004). Extracting social networks and contact information from email and the web. In First Conference on Email and Anti-Spam (CEAS).

Davison, B. (2000). Topical locality in the Web. In Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM: New York, NY, pp. 272–279.

Doan, A., Ramakrishnan, R., Chen, F., DeRose, P., Lee, Y., McCann, R., et al. (2006). Community information management. *IEEE Data Engineering Bulletin, 29*(1):64–72.

Hawking, D., & Craswell, N. (2001). Overview of the TREC-2001 web track. NIST special publication, pp. 61–67.

Heckerman, D., Meek, C., & Koller, D. (2007). Probabilistic entity-relationship models, PRMs, and plate Models. Introduction to Statistical Relational Learning.

Jordan, M. (1998). Learning in graphical models. Norwell, MA: Kluwer.

Kraaij, W., Westerveld, T., & Hiemstra, D. (2002). The importance of prior probabilities for entry page search. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 27–34). New York, NY: ACM Press.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In International Conference on Machine Learning, pp. 282–289.

McCallum, A. (2003). Efficiently inducing features of conditional random fields. In Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI03).

Murphy, K., Weiss, Y., & Jordan, M. (1999). Loopy belief propagation for approximate inference: An empirical study. In Proceedings of Uncertainty in AI, Citeseer, pp. 467–475.

Neville, J., & Jensen ,D. (2003). Collective classification with relational dependency networks. In Proceedings of the Second International Workshop on Multi-Relational Data Mining, pp. 77–91.

Ogilvie, P., & Callan, J. (2003). Combining document representations for known-item search. In Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 143–150). New York, NY: ACM.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Mateo, CA: Morgan Kaufmann Publishers.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1):1–47.

Shakes, J., Langheinrich, M., & Etzioni, O. (1997). Dynamic reference sifting: A case study in the homepage domain. *Computer Networks and ISDN Systems, 29*(8–13):1193–1204.

Tang, J., Zhang, D., & Yao, L. (2007). Social network extraction of academic researchers. In Seventh IEEE International Conference on Data Mining, pp. 292–301.

Taskar, B., Abbeel, P., & Koller, D. (2002). Discriminative probabilistic models for relational data. In Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI02), pp. 895–902.

Upstill, T. Craswell, N., & Hawking, D. (2003). Query-independent evidence in home page finding. *ACM Transactions on Information Systems (TOIS), 21*(3):286–313.

Voorhees, E., & Harman, D. (2001). Overview of TREC 2001. NIST Special Publication, pp. 500–250.

Westerveld, T., Hiemstra, D., & Kraaij, W. (2002). Retrieving web pages using content, links, URLs and anchors. NIST special publication, pp. 663–672.

Xi, W., Fox, E., Tan, R., & Shu, J. (2002). Machine learning approach for homepage finding task. Lecture Notes in Computer Science, pp. 145–159.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM, pp. 42–49.

Yang, Y., & Pedersen, J. (1997). A comparative study on feature selection in text categorization. In International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, pp. 412–420.

Yedidia, J., Freeman, W., & Weiss, Y. (2001). Generalized nelief propagation. Advances in Neural Information Processing Systems, pp. 689–695.