On Identifying Academic Homepages for Digital Libraries

Sujatha Das
Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802
gsdas@cse.psu.edu

C. Lee Giles, Prasenjit Mitra,
Cornelia Caragea
Information Science and Technology
The Pennsylvania State University
University Park, PA 16802
{giles, pmitra, ccaragea}@ist.psu.edu

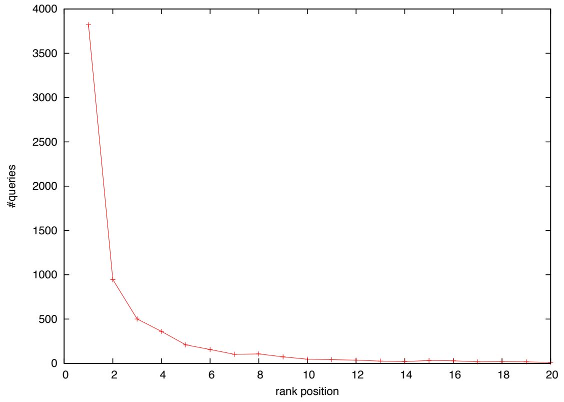
ABSTRACT

Academic homepages are rich sources of information on scientific research and researchers. Most researchers provide information about themselves and links to their research publications on their homepages. In this study, we address the following questions related to academic homepages: (1) How many academic homepages are there on the web? (2) Can we accurately discriminate between academic homepages and other webpages? and (3) What information can be extracted about researchers from their homepages? For addressing the first question, we use mark-recapture techniques commonly employed in biometrics to estimate animal population sizes. Our results indicate that academic homepages comprise a small fraction of the Web making automatic methods for discriminating them crucial. We study the performance of content-based features for classifying webpages. We propose the use of topic models for identifying content-based features for classification and show that a small set of LDA-based features out-perform term features selected using traditional techniques such as aggregate term frequencies or mutual information. Finally, we deal with the extraction of name and research interests information from an academic homepage. Term-topic associations obtained from topic models are used to design a novel, unsupervised technique to identify short segments corresponding to research interests of the researchers specified in academic homepages. We show the efficacy of our proposed methods on all the three tasks by experimentally evaluating them on multiple publicly-available datasets.

1. INTRODUCTION

Homepage finding was a well-researched task at TREC [1]. The interest in accurate identification and ranking of homepages stems from the fact that they comprise the correct answers to navigational queries which form a large proportion of queries on the Web [7] [37]. We focus on the problems related to academic homepages or professional homepages of people commonly engaged in research activities including faculty and students in universities and employees of research labs. Not only can several people with the same name have homepages, the same person can have multiple webpages associated with him/her making simple querying with the name and ranking using term match features ineffective. Figure 2 shows anecdotal examples with a couple of researcher names in our dataset. For instance, when 'Michael Jordan' the name of a famous researcher in machine learning is used as a query on Google [2], the URL we are interested in appeared at the 36th position in the search results. In this case, there are several pages related to the basketball player in the remaining results. For most researchers in Computer Science, we found that it is common to see several pages for instance, from wikipedia, DBLP [3], networking sites such as linkedin, book websites, and other institutional pages. Figure 1 shows a plot showing the rank position at which the correct homepage was found using researcher names as queries on the Web (DBLP dataset, Section 5). As the figure indicates for a reasonable number of the queried names, the correct homepages appear in positions beyond the top-3 positions of the search results. Automatic techniques for discriminating the correct homepage are therefore essential for assimilating a collection of academic homepages.

Figure 1: Rank position of Homepages in Search Results using names in the DBLP dataset (6572 queries)



In this paper, we use the phrase "academic homepage" in the following sense: the webpage of a person that has information about a person's professional status in a university

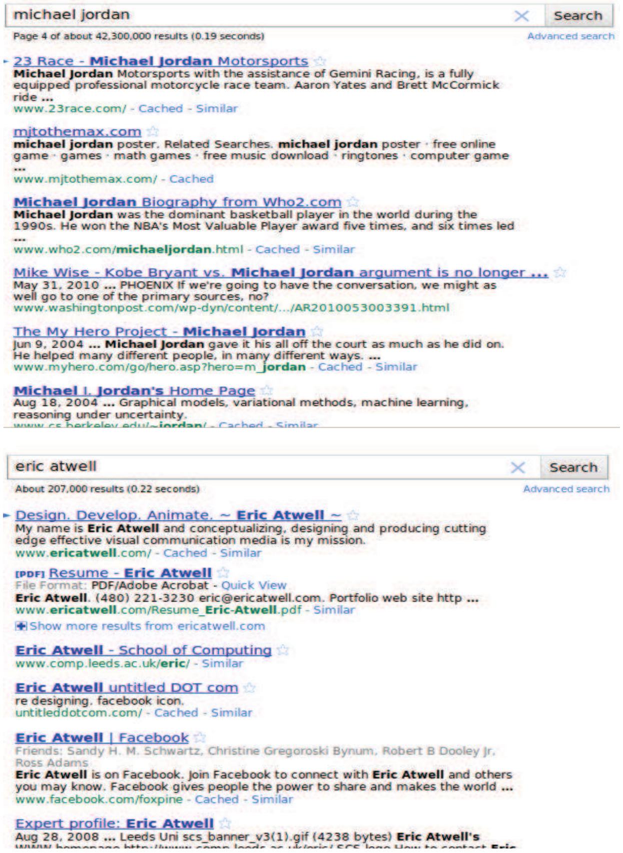
[1] http://trec.nist.gov/

[2] search performed on Oct 17th, 2010

[3] http://www.informatik.uni-trier.de/l̃ey/db/

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.

Figure 2: Example Homepage Searches



michael jordan
Search
Page 4 of about 42,300,000 results (0.19 seconds)
Advanced search

23 Race - **Michael Jordan** Motorsports
Michael Jordan Motorsports with the assistance of Gemini Racing, is a fully equipped professional motorcycle race team. Aaron Yates and Brett McCormick ride ...
www.23race.com/ - Cached - Similar

mjtothemax.com
michael jordan poster. Related Searches. **michael jordan** poster · free online game · games · math games · free music download · ringtones · computer game ...
www.mjtothemax.com/ - Cached

Michael Jordan Biography from Who2.com
Michael Jordan was the dominant basketball player in the world during the 1990s. He won the NBA's Most Valuable Player award five times, and six times led ...
www.who2.com/**michaeljordan**.html - Cached - Similar

Mike Wise - Kobe Bryant vs. **Michael Jordan** argument is no longer ...
May 31, 2010 ... PHOENIX If we're going to have the conversation, we might as well go to one of the primary sources, no?
www.washingtonpost.com/wp-dyn/content/.../AR2010053003391.html

The My Hero Project - **Michael Jordan**
Jun 9, 2004 ... **Michael Jordan** gave it his all off the court as much as he did on. He helped many different people, in many different ways. ...
www.myhero.com/go/hero.asp?hero=m_**jordan** - Cached - Similar

Michael I. Jordan's Home Page
Aug 18, 2004 ... Graphical models, variational methods, machine learning, reasoning under uncertainty.
www.cs.berkeley.edu/~**jordan**/ - Cached - Similar

eric atwell
Search
About 207,000 results (0.22 seconds)
Advanced search

Design. Develop. Animate. ~ **Eric Atwell** ~
My name is **Eric Atwell** and conceptualizing, designing and producing cutting edge effective visual communication media is my mission.
www.**ericatwell**.com/ - Cached - Similar

[PDF] Resume - **Eric Atwell**
File Format: PDF/Adobe Acrobat - Quick View
Eric Atwell. (480) 221-3230 eric@ericatwell.com. Portfolio web site http ...
www.**ericatwell**.com/Resume_**Eric-Atwell**.pdf - Similar
Show more results from ericatwell.com

Eric Atwell - School of Computing
www.comp.leeds.ac.uk/**eric**/ - Similar

Eric Atwell untitled DOT com
re designing. facebook icon.
untitleddotcom.com/ - Cached - Similar

Eric Atwell | Facebook
Friends: Sandy H. M. Schwartz, Christine Gregoroski Bynum, Robert B Dooley Jr, Ross Adams
Eric Atwell is on Facebook. Join Facebook to connect with **Eric Atwell** and others you may know. Facebook gives people the power to share and makes the world ...
www.facebook.com/foxpine - Cached - Similar

Expert profile: **Eric Atwell**
Aug 28, 2008 ... Leeds Uni scs_banner_v3(1).gif (4238 bytes) **Eric Atwell's** WWW homepage http://www.comp.leeds.ac.uk/eric/ SCS logo How to contact **Eric**

or a research institution indicating his or her research interests, projects and listing out or linking to his or her publications. Academic homepages serve as concise descriptions of a person's academic standing and current activity and are a valuable resource for digital libraries like CiteSeerX [22] and ArnetMiner [36]. These portals provide access to collections of research literature and information regarding their associated authors. Backend tasks of such portals include crawling for new literature, metadata extraction from research documents and document classification. Academic homepages of authors serve as a resource that can help in several of these tasks. For instance, a crawler can potentially obtain up-to-date publications related to authors from their homepages. Similarly, metadata available from homepages such as e-mail information or affiliation can help in disambiguating author names [18].

1.1 Our Contributions

In the rest of the paper, we refer to "an academic homepage" using just the term 'homepage'. We try to answer the following questions in this study: (1) How many academic homepages are there on the web? (2) Can we accurately discriminate between academic homepages and other webpages? and (3) What information can be extracted about researchers from their homepages?

For the first question, we use mark-recapture, a method widely used to estimate animal population sizes in Biometrics [28]. Our experiment with capture samples of homepages in the Computer Science domain provides experimental evidence to our guess that homepages form a small fraction of the Web. Nevertheless, given their resourcefulness in digital library tasks, it is very desirable to accurately discriminate academic homepages from the "rest of the web". This prompts us to address our second question related to homepage identification in the context of crawling. As opposed to the TREC task or previous work on academic homepage finding (such as in ArnetMiner [36]), the set of names of the persons whose homepages we are interested in, is usually not available apriori in this situation. Therefore, we study the content-specific aspects of a homepage with the goal of identifying homepages in absence of name information. Based on our analysis with generative topic models (Latent Dirichlet Allocation or LDA) on known homepages, we posit that an average academic homepage can be viewed as a mixture of "categories of information" (or topics in LDA). We study the performance of various features based on the topic distribution vector of homepages and the words from discriminative topics on the homepage classification task. Experimental results are provided on publicly available datasets from three different sources using both sets of features. Our experiments demonstrate that academic homepage identification on the Web is rather challenging but content-based features can serve as an effective filtering mechanism.

Digital libraries provide an effective resource for solving interesting problems such as expertise ranking and mining influential authors [25, 10]. Authors are represented in digital libraries in terms of their expertise areas and other metadata such as affiliation and contact data. Tang, et al [35, 36] showed that researcher homepages are an effective source for extracting metadata information. We focus on extracting the name of the researcher and his or her research interests from his or her homepage as part of answering our third and final question. We employ existing work in named-entity recognition for name extraction and propose a simple technique using the term-topic associations obtained from known homepages to extract the "research interests" segment from an academic homepage. This unsupervised technique is novel in the face of most existing approaches to metadata extraction that involve supervised learning.

The rest of this paper is organized as follows: Previous research related to our contributions is briefly summarized in Section 2. Details on our techniques are provided in Sections 3 and 4 whereas Section 5 covers the details of evaluation and observations. We conclude with a summary and future directions for our work.

2. RELATED WORK

Webpage Classification and Homepage Finding

Due to the overwhelming size of the Web, techniques for automatic webpage classification are well-studied. Webpage classification is addressed from at least two perspectives–topic and genre. Qi and Davison provide an elaborate survey on the approaches for topic-oriented webpage classification [31]. Here, the target classes could be related to subject (E.g. arts or sports), sentiment (E.g. opinion pages) or function (E.g. coursepage or homepage). Content-based features are commonly used for subject-based classification in contrast to genre identification. Genres such as FAQs, blogs, e-shops and news are defined based on the target audiences for these pages. Visual and structural aspects of HTML were found to be effective for genre identification [23, 19].

Several researchers employed machine learning techniques to address the homepage finding task at TREC [40, 37, 27]. Although the focus was not on academic homepages in particular, the general observation from research in this area is that query-independent features from the content, anchor-text, URL-type and PageRank can be combined with query-dependent features to significantly improve classification and ranking performance. Tang, et al. identified academic homepages using features based on queries in ArnetMiner [36]. In response to a person name query, they used binary SVM classifiers trained with features such as the presence of the person's name in the HTML title and the URL of pages retrieved via the Google API.

In contrast with the TREC task and ArnetMiner, we seek to crawl the web and identify homepages for adding to digital libraries. This scenario is similar to that of Wang and Oyama [38] who target building a high-quality collection of researchers' homepages. Wang and Oyama combine keywords on a webpage along with those on local surrounding pages for academic webpage identification. They manually identified several keywords in Japanese, commonly found in homepages and categorized them into "property" lists such as title (doctor, professor), major (major/research field) etc. In our work, we focus on automatic means to derive such lists using generative topic models. In particular, we use Latent Dirichlet Allocation (LDA) [5, 14, 17] for deriving features for classification.

Population Estimation and Expert Profiling
Mark-recapture techniques are widely studied and applied in Biometrics for estimating population sizes of animals [28, 29, 20]. These models were also recently used to study the size of the Web [21, 12, 4]. In particular, Gibbs sampling [13] was used to estimate the size of the telephone universe [30] and the number of robots on the Web [34]. We show an estimation experiment using mark-recapture and Gibbs sampling for calculating the number of academic pages in the Computer Science domain on the Web.

Balog and Rijke designated the "record of types and areas of skills" of a person as the "topical profile" for that individual [1]. Usually profiling involves modeling a candidate's profile using the documents associated with that individual. Instead, our focus is on obtaining the author's (self-described) topical profile from his or her homepage. Tang, et al. used homepages to generate researcher profiles in ArnetMiner [36]. With the goal of integrating researchers' personal information into a digital libraries Tang, et al. define an elaborate scheme for a profile including attributes such as name, photo, affiliation, interests, email, etc. Conditional Random Fields were employed to learn a tagger for this schema using various content and pattern based features. This approach is very similar to most other approaches to metadata extraction from unstructured or semi-structured text. Other examples of such tasks include extracting named entities from web pages [39] and author and title extraction from academic papers [16]. Zheng, et al. extract author metadata information from homepages using visual features [42]. Again, they used supervised machine learning techniques to identify segments in the homepage corresponding to author name, affiliation, picture, etc. Their ontology does not include an author's "research interests", a rather significant field for digital libraries. As noted by these authors, unsupervised methods such as template detection and wrapper induction [41] commonly used for extracting metadata from product websites like Amazon are unlikely to work for academic homepages. Researchers style their homepages based on their own preferences and very rarely use institute-provided templates (if available) for this purpose. Even if they did, these templates differ from institute to institute making the design of a generic wrapper impractical. We focus on the extraction of name and "research interests" segments from an academic homepage. We propose heuristic techniques for extracting author names and "research interests" from researcher homepages. Despite being unsupervised, we show that our techniques perform quite well and do not require explicit training on manually annotated datasets.

3. ESTIMATING THE NUMBER OF HOMEPAGES ON THE WEB

Mark-recapture methods are probabilistic techniques employed in biometric studies for estimating population sizes of birds and animals in a certain area [28]. These techniques involve obtaining samples from the population of interest and counting the number of individuals that appear multiple times in the collected samples. Let the population consist of N (unknown) individuals and suppose that a sample of n_1 individuals was captured the first time. These individuals are marked and released into the population. Some time later a second sample of size n_2 is caught. Let m be the individuals which were seen in the first sample also seen in the second sample (identifiable since they were marked and released back, sampling with replacement). Under a closed-world assumption (N did not change during the sampling process) and assuming the capture probabilities are the same ($\frac{n_1}{N} = \frac{m}{n_2}$) Lincoln-Peterson method [28] estimates the population $\hat{N} = \frac{n_1 n_2}{m}$. More generally, let I be the number of samples collected and the probability that an individual j is captured in sample i be p_{ij}. In a homogeneous catch model, the probability of capture is assumed to identical for all individuals in a sample, that is $p_{ij} = p_i$. Let $n_1 \dots n_I$ be the sizes of samples drawn, marked and returned to the population and the total number of distinct captured individuals be r. The likelihood function of N and $p = (p_1, ...p_I)$ from data D is given by

$$L(N, p|D) \propto \frac{N!}{(N-r)!} \prod_{i=1}^{I} {p_i}^{n_i} (1-p_i)^{N-n_i}$$

George and Robert [13] show that with appropriate prior distributions on N and p, conditional posterior distributions for N and p that are easy to sample from can be derived and estimates of N obtained via Gibbs Sampling. For instance, assuming Jeffreys prior $\pi(N) = 1/N$, the conditional posterior of N is negative binomial with parameters $r-1$ and $1 - \prod(1-p_i)$ and with independent $Beta(a, b)$ priors on p_i's, the conditional posteriors of p_is are independent $Beta(n_i + a, N - n_i + b)$. In addition to animal population estimatations, Gibbs sampling was used to estimate sizes of other types of population such as the Web [12], the telephone universe [30] and more recently the number of robots on the Web [34]. We use this estimation process in Section 5 to count the number of academic homepages in Computer Science (CS) on the Web using our datasets as "captured" samples. To our knowledge, we are the first to use mark-recapture estimation to count the number of pages on the web that belong to a certain type (academic homepages). Broder, et al. proposed techniques for estimating a corpus size using the query interface of a search engine [8]. In contrast, name (or query) information is not available to us apriori and indeed extracting this information is one of the objectives in this paper.

4. CHARACTERISTICS OF HOMEPAGES

People in general and particularly academic scholars create homepages to indicate their presence on the Web and to advertise their work. The following observations are based on analyzing samples of homepages of the type we are interested in, that is, those that are useful from the perspective of a digital library. It appears that a researcher's homepage comprises of certain **specific categories of information**. For instance, in the Computer Science domain an academic homepage usually contains the person's affiliation and contact information in addition to a brief summary of the person's background and current academic activities. For example, a faculty member usually indicates her research activities, her membership status in various committees, her teaching activities and so on, on her homepage. It is also common to see a list of publications or a link to the same on such a homepage.

Table 1: Top words from Topics of Homepages

talk	page	students	member
slides	home	graduate	program
invited	publications	faculty	committee
part	links	research	chair
talks	contact	cse	teaching
tutorial	personal	student	board
seminar	list	undergraduate	editor
summer	updated	college	courses
book	fax	current	state
introduction	email	ph	activities
chapter	department	school	technical
group	interests	program	associate
workshop	phone	university	special
lectures	info	grant	education
presentation	homepage	news	present

4.1 Homepages as topic mixtures

Latent topic mixture models posit that a document can be viewed as a mixture of a small number of latent topics and that each 'observed' word in the document can be attributed to one of these topics. Note that the process of homepage creation by its author can be visualized in a fashion similar to that of the document generation process inside topic models such as LDA (Latent Dirichlet Allocation). If each "category of information" is mapped to a topic, the creator of a homepage seems to be adopting the following steps while generating his or her homepage.

1. Sample a mixture proportion of topics. Each topic corresponds to a specific category of information such as contact information, publications etc. Depending on the person's preference, the homepage might contain more information related to one category (topic) than the others. Similarly, the layout or position of each category of information varies depending on personal preferences.

2. For each of the N terms in the homepage
   (a) Sample a topic for that position
   (b) Sample a word conditioned on the chosen topic

To avoid clutter and focus on the intuition, we deliberately skipped the mathematical notation of LDA in the above description. The details of LDA including the plate notation, sampling equations and the estimation process can be found in the references [5, 14, 17]. Previous research using LDA has shown its effectiveness as an unsupervised tool for analyzing text corpora. We now describe some quantities that LDA estimates from a collection of documents since they are used in later sections for classification and profile extraction. Based on co-occurrence counts in the corpus, LDA learns a topic-term association matrix, ϕ. The entries in this matrix corresponds to predictive distributions of words given topics, that is, $\phi_{w,i}$ is the probability of a word w given the topic i. After an LDA run, every term in the document is randomly assigned a topic based on these probabilities. These assignments are used to express the document as a mixture of topics. θ_d refers to the topic distribution vector of length K for document d, where K is the number of topics (a parameter while running LDA). The component $\theta_{d,i}$ is the smoothed proportion of times topic i was assigned to the terms in d. From an analysis standpoint, obtaining the top words for a given topic (words with high probability values for that topic from ϕ) usually helps in discerning the underlying theme captured by that topic. Table 1 shows the top words of topics indicative of homepages obtained by running LDA on known homepages in our DBLP dataset (Section 5.1). Note how these topics capture the "categories" of information expressed by authors in homepages that we described earlier. Note that, in addition to content, homepages tend to follow certain structural conventions. For instance, it is very common for homepages to be hosted on the university or the institute domain the researcher is affiliated with. In addition, homepage URLs that belong to a certain university usually follow a particular naming convention for the URL (For example, a tilde followed by author's lastname after the department URL). While URL features might be less consistent across institutions, certain HTML features might be common among homepages. For instance, it is a common convention to put the author's name, sometimes coupled with the word 'home' in the title tag of the HTML. Similarly, it is fairly uncommon for academic homepages to contain several tables or images embedded in them. We leave the design of structural features for homepage classification for future work and focus on content-based approaches in this paper.

4.2 List of Feature Sets

We tested the following feature sets in our experiments:

1. **All Topic Proportions (ATP)**: The components of the topic distribution vector, θ_d output by LDA for a given document are used as features for that document.

2. **Specific Topic Proportions (STP)**: We use greedy feature selection to identify among all topics output by LDA a subset of topics that is indicative of homepages (next subsection). Only topic proportions related to

Table 2: Top words from topics on subject areas

data	multimedia	systems	design
database	content	distributed	circuits
databases	presentation	computing	systems
information	document	peer	digital
management	media	operating	signal
query	data	grid	vlsi
systems	documents	storage	ieee
xml	based	middleware	hardware
acm	hypermedia	system	fpga
vldb	video	scale	implementation
sigmod	user	high	power
icde	adaptation	large	architectures

this subset are used as feature values in classification. This set is designed to avoid learning a classifier that uses topical features not related to homepages. For instance, LDA also identifies topic clusters that indicate subject areas which we use for other purposes (Section 4.4).

3. **Word Features (WF)**: The list of words based on the top words of topics indicative of homepages were used as features. These topics were manually identified from the output of LDA. The feature vector for each document comprises the normalized term frequencies of the words in this list. Using the words of specific topics as features provides a finer granularity instead of an aggregate topic proportion value.

In Section 5, we check the performance of various classifiers with the above sets of features. As shown in the results section, the classification performance is highly dependent not only on the set of features but also the classification algorithm used.

4.3 Identifying Topics Indicating Homepages

Exploratory analysis with LDA involves running the model with a specific number of topics on a large collection of documents, estimating various parameters of the model and manually examining the output of the model. As mentioned previously, the output from LDA includes clusters of terms that are highly probable for each topic. These words usually indicate the underlying theme covered by the given topic. For instance, the top words shown in Table 1 are indicative of themes like "contact information" (second column) and "professional activity" (last column). However, when LDA is run on homepages, the topics identified do not necessarily indicate homepage-like aspects. For instance, some of the topics identified with LDA on our dataset pertain to *subject areas* and are shown in Table 2. These topics extracted by LDA are evidence that it is common to find information related to "research interests" or "area of work" on homepages. However, the exact nature of these topics is more an artifact of the dataset. Our dataset was based on homepages from Computer Science and the words in the table clearly show topics corresponding to databases, multimedia, distributed systems and circuit design, the subfields in Computer Science. A classifier for identifying homepages if trained using these features would make our method domain-dependent. We therefore need means to automatically identify topics which help homepage classification but are domain-independent at the same time. The super-greedy feature selection algorithm was found to be effective in identifying the required subset of topics [11]. This simple technique involves sorting the features by their LOOCV (leave-one-out-cross-validation) and taking the top few features as the selected subset. We used the performance on the validation set as a measure while selecting the subset. That is, we run this algorithm by using word features for each topic and retain only the top topics that contribute most to the F1 score on the validation set. Forward selection, a commonly used strategy for feature selection, that evaluates subsets of features tended to select larger set of topics than actually required for a given F1 score.

4.4 Extracting Name and Research Interests

As opposed to the general problem addressed in ArnetMiner, we focus on extracting the name and research interests segments from a researcher's homepage. After crawling a homepage it is desirable to map it to an existing entity or create the placeholder for a new entity in the homepage collection. Name information inside the homepage can serve this purpose. Other information such as e-mail and affliation can further help in diambiguation in case of multiple people with the same name. For the purpose of expert rank and profiling [1] name information and research interests coupled with publications associated with the researcher arguably play a major role. Some previous attempts to extract research profiles from homepages were discussed in related work. As opposed to these supervised methods, in this section, we investigate if the subject-area topics identified by LDA (Table 2) can guide the extraction of the research interests segment from a homepage in an unsupervised manner. We define segment as a short consecutive sequence of words (usually $20 - 30$ words long) inside a homepage.

As described earlier, after an LDA run each term in a document is assigned a topic based on the topic-term associations matrix, ϕ and these assignments enable the expression of a document in terms of its topic mixture or the topic proportion vector (θ_d). Furthermore, as we observed earlier, it is very common for researchers to mention their areas of interests (subject areas) on their homepages manifest in the subject-specific topics identified by LDA. It is likely that the words in segments of the homepage that pertain to research interests are assigned topics related to subject areas (of the researcher). However, assignments from LDA do not necessarily ensure that sequences of words describing the same topic are indeed assigned the exact same topic. Although this is likely, the random sampling process in LDA does not ensure this. Suppose t is the topic related to the research interest of the researcher and w is a word inside the research interest segment. While w may not be assigned t as part of the final LDA assignments, it is likely that the term-topic association value $\phi_{w,t}$ is high for this pair. We can use this idea to describe score for a given segment (s) inside the homepage with respect to a topic t as $Score(s,t) = \sum_{w \in s} \phi_{w,t}$.

This score now permits ranking of segments in the homepage for a given topic. The topical profile or "research interests" segment of the homepage is designated as:

$$p = argmax_{t \in ST, s \in S} Score(s,t)$$

Here S is all possible segments in the homepage with a given size sz and ST is the set of all topics indicating subject areas. The profile segment therefore comprises the words in the segment indicated by p. Note that researchers usually have multiple (possibly related) research areas and our current scoring function needs to be extended to handle this case. For example, this can be done by considering subsets of subject-topics instead of single topics while computing the score.

We used named-entity extraction features for identifying extracting researcher names from their homepages. Named-entity extraction is a well-known research problem in natural language processing (NLP) where the goal is to identify names of entities such as persons, companies, locations and organizations from free text [26]. We use the following heuristic for handling the name extraction from homepages. The first "person name" that appears on the homepage is most likely that of the person whom the homepage is about. We demonstrate in Section 5 that this simple heuristic is quite successful in extracting the researchers' name information from their homepages. However, it is clear that this heuristic depends on the performance of the named-entity extraction software which is itself usually a tagging algorithm trained using supervised techniques and involves several language-dependent features such as punctuation hints, capitalization, gazetteers and so on [26].

5. EXPERIMENTS

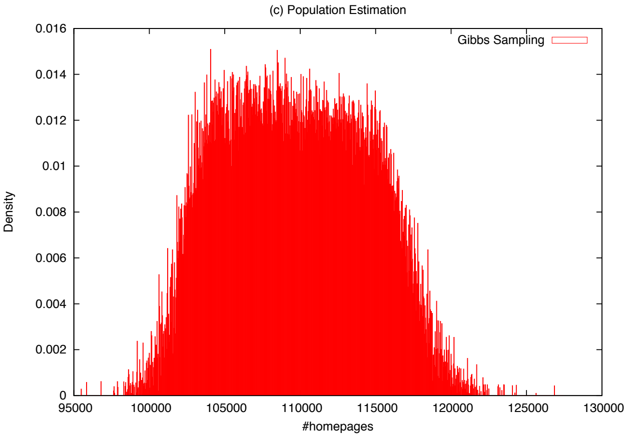
5.1 Datasets

In context of crawling, our classifier can be considered competitive if it can discriminate between academic vs. non-academic webpages, homepages among other academic pages such as course or department pages and homepages of researchers in other subject domains. We consider the following datasets for evaluating our classifier.

1. **DBLP Dataset** This is a homegrown dataset created by obtaining the author names from DBLP in January 2010. DBLP provides bibliographic information related to computer science journals and proceedings in areas like databases, networks, machine learning and so on. DBLP indexes more than one million articles and names of computer scientists whose articles are listed there. At the time of our dataset creation, DBLP listed about 769785 author names out of which for 13290 authors, the homepage url information was also available. Each author name for which the homepage URL was specified was used as a query string to search the web with Yahoo's BOSS API [4]. The first 20 hits from this search were scanned for the homepage URL listed in DBLP. If the homepage was found in the top-20 hits, this was marked as a positive example and the remaining 19 hits comprise the negative examples. As can be seen, this process results in a highly unbalanced dataset but since our objective is to make the homepage crawler tenable in similar situations on the Web, this unbalanced dataset appears to be a good one to train on.

2. **WebKB** The WebKB dataset [5] contains about $8,282$ webpages from universities categorized into student, faculty, staff, department, course, project and other pages. The pages categorized under student and faculty are treated as academic homepages whereas other pages comprise the negative examples. The pages in the 'other' category were ignored since they represent pages that branch off pages in the remaining categories and could be linked to homepages.

Figure 3: Population Estimation with Gibbs Sampling



3. **Eprints** The E-print network [6] comprises a valuable resource of homepages and publication pages of researchers in different disciplines including Environmental Sciences, Chemistry etc. Since the name information is not directly available, we could not obtain negative examples as we did with the DBLP dataset. The homepages in this set, particularly in other subject domains serve as good testsets to evaluate the "domain-independent" nature of our features.

The number of instances for which both structural as well as content-based features could be extracted for each of the above datasets are summarized in Table 3.

Table 3: Dataset Statistics

Dataset	NumPositive	NumNegative
DBLP	6367	112362
WebKB	1772	2945
Eprints-All	9359	-

5.2 Estimation Results and Observations

The three datasets (DBLP, WebKB and Eprints) can be treated as "capture" samples for estimating the population of academic homepages in Computer Science (CS) domain. For these captures, the value of r was found to be 19177 and values of $\{n_1, n_2, n_3\} = \{13290, 2764, 4035\}$. Applying Lincoln-Peterson's formula taking two of these datasets at a time gives crude estimates of homepage population size as 79563, 170850 and 259370. We also implemented Gibbs Sampling as discussed in Section 3 using these capture samples. Assuming the samples to be independent (as far as we know these sources are independent from each other) and $Beta(1,1)$ priors for the p_i distributions, we used SSJ [7], the stochastic simulation library in Java for sampling. Figure 3 shows the density vs. population histogram plot for the collected samples. Every 20^{th} sample from a run of 2000000 iterations was recorded after a burn-in of 10000. A point-estimate using these samples gives the population size of

[4] http://developer.yahoo.com/search/boss/

[5] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

[6] http://www.osti.gov/eprints/

[7] http://www.iro.umontreal.ca/s̃imardr/ssj/indexe.html

homepages in CS as 109551. Note a few caveats of the estimation process. First, a Beta(1,1) prior on capture probabilities is essentially a non-informative one. The fact that DBLP lists about 769785 authors but only about 13290 homepages, potentially hints at a prior other than Beta(1,1). However, there does not seem to be an obvious method to set priors in our case (such as using cross-validation commonly employed in bayesian classification). The calculated estimate is known to display large variation depending on the chosen priors [13]. Obtaining unbiased and independent samples on the web is also known to be difficult task [4, 3, 2]. The objective in this experiment was to illustrate that mark-recapture methods can be used for estimating the population sizes of specific types of pages. The estimate becomes more accurate with more capture samples, better priors and with a more representative model. For instance, it appears that the probability of researchers having a homepage is different for different subject areas or in other words, not every researcher is equally likely to have a homepage. Such extra information canbe used to derive a mixture model similar to that used by Poole [30]. In this model, a set of groups is defined each having different capture probabilities and individuals have probabilities of belonging to a particular group. Even if our estimate is off by a multiple or even an order of magnitude, this number would still be considerably small compared to the rest of the indexed web (around 2.7 billion on Oct. 29, 2010 [8]). Nevertheless, given the richness of information present in them, efficient techniques for accurately identifying them become even more crucial.

5.3 Classification Results

We primarily trained our classifier on the DBLP dataset since the instances here are more general than that of WebKB and negative instances were unavailable for Eprints. The author name information is available for this set enabling comparison with query dependent features. Three random splits of 70/30 proportions were created from this dataset. The results shown on DBLP datasets are averaged results across all splits. We first summarize our experiments with this dataset. We used the LDA implementation provided with Mallet [24] on the positive instances in our training data. The LDA estimation process requires setting the number of topics (K) value and other parameters for priors of hyperparameters. We used the default settings along with the hyperparameter optimization option available in Mallet. For number of topics, we experimented with settings 10-200 in steps of 10 and evaluated the training data likelihood. We found the likelihood value with $K = 70$ to be among the better ones. Manually examining the top words for each topic also indicated that for $K > 70$, several words are repeated under different topics indicating that multiple topics might be covering the same theme. Therefore K was set to 70 in all LDA related experiments. Typically only about 10% of the identified topics (about $5 - 7$) correspond to homepages. For word features, we only used the top-20 words from homepage related topics. Typically, this results in less than 100 words. Using larger values for topwords did not help in terms of classification. We used the normalized term counts of these words in documents as feature representation. This was found to perform better than using boolean features or raw term counts.

Table 4: Classification results with LDA-based features

Features	Precision	Recall	F1
	Decision Tree		
ATP	0.3181	0.2381	0.2719
STP	0.4498	0.0785	0.1279
WF	**0.4721**	**0.3627**	**0.4102**
	Logistic		
ATP	0.4937	0.1059	0.1735
STP	0.4348	0.0726	0.1235
WF	0.5683	0.2160	0.3130
	One-class SVM		
ATP	0.1477	*0.5354*	0.2283
STP	0.1756	*0.5227*	0.2611
WF	0.2992	*0.5027*	0.3751

Table 4 shows the performance of the feature sets described in Section 4 with various classifiers. We use precision ($\frac{tp}{tp+fp}$), recall ($\frac{tp}{tp+fn}$) and the F1 score ($\frac{2*precision*recall}{precision+recall}$) measures used commonly in evaluating classifiers. Here tp is the number of true positives, fp the number of false positives and fn is the number of false negatives. Although $F1$ score gives an aggregate measure combining precision and recall, in context of crawling recall might be a more appropriate measure. If future filtering and processing steps in digital libraries are capable of throwing out irrelevant pages, obtaining a larger fraction of the relevant pages among those present is more important.

Classification performance using all topic proportion (ATP) values or specific topic proportion (STP) values are worse that using words of specific topics (WF). Including word features rather than the coarser value of the topic proportion improved classification. The ATP setting seems to be better than the STP setting. However, using ATP is undesirable, since as we mentioned before, this includes domain-specific topics which is an artifact of the dataset we use for training. Given that the set of word features is small (around 100), we found decision trees to be the best performing among those we tried with respect to the F1 measure. Two-class SVMs with default settings were found to be very sensitive to the unbalancedness of the dataset. Neither the weighted variant (that biases learning the positive class) or a balanced version where negative examples equal in number to the positive examples were randomly sampled outperformed the classification results presented in the table (F1 around 0.05). Logistic regression was found to be robust with respect to the unbalancedness but its performance is not as good as the other classifiers. We also experimented with one-class SVMs on this task.

One-class SVMs were originally proposed for novelty detection or outlier identification [32]. Unlike 2-class or multiclass SVMs that learn separating hyperplanes between classes during training, one-class SVMs are designed to learn the representation for a single (positive) class in terms of a hypersphere and anything falling outside the hypersphere is considered an outlier or belonging to the negative class. Note that in our situation, while academic homepages comprise the positive class, there is a lot of variety in the negative class. Indeed, a negative instance could be another academic page such as a course, a page from Amazon or DBLP or something else. It is in cases like this where learning a representation for the negative class is difficult due to the underlying diversity and where identifying the positive class is of more interest, that one-class SVMs are found to

[8] http://www.worldwidewebsize.com/

be useful. We used the LibSVM [9] implementation with the c parameter set to 0.01 for one-class SVMs whereas classifier implementations in Weka [15] were used for the remaining cases. One-class SVMs are more scalable as the number of features increase compared to decision trees. They also seem to have the best performance with respect to recall when compared to the other classifiers. In summary, the classification performance depends both on the set of features and the classifier used.

In order to compare word features identified by LDA with other content-based approaches, we performed a comparison experiment using words selected using the following strategies.

1. **Term Counts (Unigrams)**: Top words based on their aggregate term counts in the positive instances of the training data are chosen as features. Terms chosen using other measures such as IDF and TFIDF performed worse than those chosen based on aggregate term counts.

2. **Mutual Information (MI)**: In this strategy, top words from the positive instances of the training set when ranked by their mutual information value [9] are chosen as features.

3. **Feature Abstraction (Abs)**: Feature abstraction methods [33] are used to reduce the classifier input size by grouping "similar" features to generate *abstract features or abstractions*. Silvescu et al. showed that *abstractions* reduce the model input size and helps improve the statistical estimates of complex models (especially when data are sparse) by reducing the number of parameters to be estimated from data.

Figure 4 shows the F1 variation with number of features chosen by each strategy with one-class SVMs(a) and decision trees (b). In our experiments, we found that decision trees took considerably long time to train when the number of features increased beyond 500 as opposed to one-class SVMs which were scalable in the face of thousands of features. Although the performance of LDA-based features as shown in Table 4 is not very high, as figure 4 indicates, a very small set of features (about $100-200$) perform on par with thousands of unigram features selected based on term counts and out-perform terms selected based on mutual information and abstraction features. As more unigram features are added, the performance gets close to that of LDA possibly due to the potential overlap as we include more terms. However, unigrams selected based on term counts are not desirable since they would potentially include several domain dependent terms apart from those specific to homepages.

For the DBLP datasets, we did not include accuracy values. These were mostly between $90-95\%$ but these numbers are largely due to getting most of the negative examples right which is very large in this dataset. We present the identification accuracies on the Eprints datasets in Table 5 using decision trees. Subject area information is available in this dataset enabling us to test the domain-independent nature of our features. The identification accuracies are somewhat low for the Eprints datasets. Moreover, the classification accuracies are not uniform across all subject areas. On the WebKB dataset, we obtained a precision of 0.8137, recall of 0.3081 and an accuracy value of 0.5413. These values are rather low. Typically, classification accuracies averaged over all the six classes are published with WebKB and are usually in the $70-90\%$ range depending on the choice of features. For instance, Boulis, et al. showed cross-validation accuracies around 90% by using tens of thousands of unigram and bigram features [6]. In comparison our classes are different (homepages vs. non-homepages) and the size of our feature set is very small. We are working on an error analysis study to find out the difficulty in distinguishing homepages among the remaining types of academic pages and the distinctions among subject domains for addressing the performance on the WebKB and the E-prints datasets.

Table 5: Identification Accuracy on Eprints datasets

Domain	Word Features
Chemistry(135)	0.1852
CS(2655)	0.3657
Physics(458)	0.2402
GeoSciences(859)	0.1920
Mathematics(4082)	0.2665
EnvSciences(1170)	0.1897

5.4 Results of Name Extraction

We tested our heuristic that given a homepage, the first 'person' name in the page is the name of the person associated with the homepage. The positive instances in the DBLP dataset was used for this purpose since author names are available for this set. Stanford's state-of-the-art named-entity recognition (NER) tool [10] was used to identify person names. The performance of our heuristic, along with some anecdotal examples of "incorrect" extractions are presented in Table 6. From examination of "failed" cases, we found that most extractions got the name right as specified in the homepage although variants of these names were specified as "correct" names in the DBLP dataset. This is obvious from the rather high Jaccard similarity score although the number of exact matches is only about 30%. Computing Jaccard's coefficient involves treating the words in the extracted and correct name sequences as sets A and B and measuring the overlap using the value of $\frac{|A \cap B|}{|A \cup B|}$. Our text extractor [11] extracts the text from HTML title and places this just ahead of the text in the body of the HTML page. Since authors' tend to put their name both in the title as well as at the beginning of their homepage, we found a large number of extractions such as the "Dimitris Papadias" case in the table. The NER tool failed to extract person names for about 7% of the homepages. These results look rather good and demonstrate a quick way based on intuition to extract author names from a newly acquired homepage. However, a word of caution is needed here. The performance of this heuristic is entirely dependent on the performance of the named-entity tagger the choice of which must be made depending on the language, its performance and scalability.

5.5 Extracting "Research Interests"

We could not provide a quantitive evaluation for our research segment extraction algorithm. To our knowledge no publicly available dataset is available to evaluate this extraction. The datasets available from previous work related to extracting metadata from homepages [42, 36] do not ex-

[9] http://nlp.stanford.edu/IR-book/html/htmledition/mutual-information-1.html

[10] http://nlp.stanford.edu/ner/index.shtml

[11] http://htmlparser.sourceforge.net/

Figure 4: F1 vs #Features (a) Decision Trees & (b) One-class SVMs

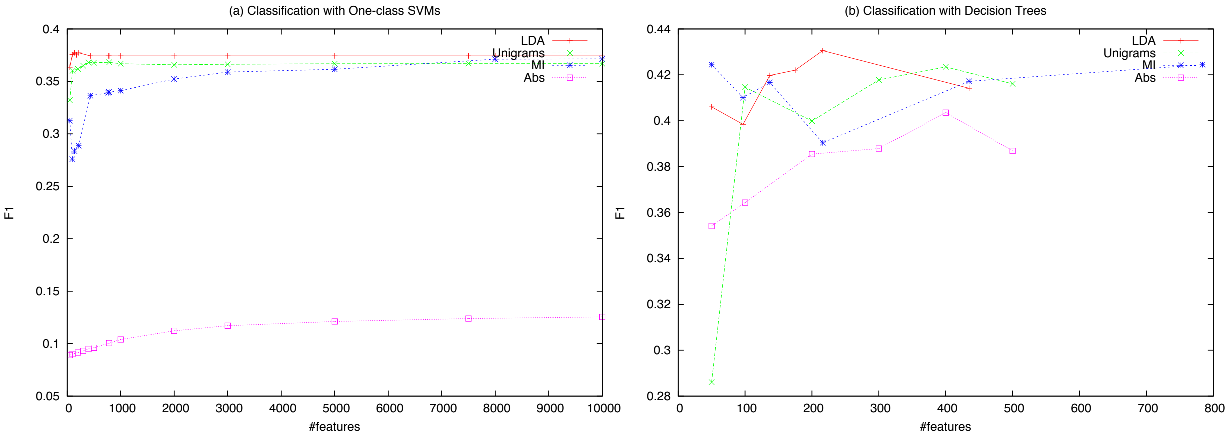


Table 6: Name Extraction on the DBLP Set

NumNames	6572
NER misses	445
Exact Matches	1985
Jaccard Sim	0.8025
Specified(dataset)	**Extracted(homepage)**
Mario Gerla	Dr. Gerla
Brian R. von Konsky	Brian von Konsky
Matthias Dehmer	Matthias Dehmer Short Vita
Dimitris Papadias	Dimitris Papadias Dimitris Papadias
Irek Ulidowski	Irek Ulidowski B. Sc
Subhash Suri	Santa Barbara

plictly annotate these fields. We are currently working on acquiring such a dataset. In the mean time, for the sake of concreteness, we provide anecdotes of our extraction algorithm on randomly selected examples from the DBLP dataset since our LDA was also trained with this dataset. The segment size was set to 20 words in these experiments. We manually selected the topics denoting subject-areas from the topics identified by LDA. A few homepage URLs and the research interests segment extracted from the text at these URLs (when we obtained them) with our algorithm are presented in Figure 5. It can be seen that our technique is rather effective in approximately identifying the research interests segment from homepages. In pair 2 of the figure, a publication page was erroneously specified as the homepage. In such instances and also in instances where descriptions of publications or projects are included in the homepage, it is likely that our algorithm makes mistakes. The extraction algorithm can be clearly improved by adding some supervision, e.g. discounting the scores of segments closer to the term "publications" or boosting the scores of segments closer to the phrase "research interests". Still, from anecdotal evidence, this unsupervised technique works rather well for a first-cut.

6. SUMMARY AND FUTURE WORK

We studied problems related to academic homepages. We used mark-recapture methods to estimate the number of academic homepages on the Web. Our estimate of homepages in the computer science domain indicates that academic homepages constitute a small fraction of the Web; however, they are a significant resource of research literature and information on researchers. Hence their collection is very desirable. We experimented with content-based approaches for classifying academic homepages. We showed that even in the absence of queries or person names, academic homepages can be identified using content-based features. We posited that homepages can be viewed as mixtures of certain categories of information. This view enabled us to apply generative topic models (LDA) and later identify "word features" for discriminating homepages. We found that a small set of features thus identified are better at classifying homepages than choosing words based on a few other feature selection strategies. We showed classification results with various classifiers on diverse datasets available for this task. Next, we presented techniques for extracting researcher names and research interests from their homepages.

We are currently working on folding in our classifier module into a web-scale crawler. University faculty lists form the seeds for such a crawl. There is still room for improvement with respect to the classification measures on all the datasets. Analysis of the specific nature of homepages in the failed cases is a necessary first-step. In addition, our preliminary experiments with HTML-based features (not included in this paper) indicate that homepages also show certain properties with respect to their structural layout and URL strings. For example, homepages rarely contain too many tables or images. We hope to augment content-based features with features designed based on these observations for a more accurate identification of homepages. Similarly, adding more capture samples (and across domains) can provide a more accurate size estimate for the number of homepages. The final goal is the extraction of metadata and publication information from academic homepages. We are currently exploring whether LDA estimates can enable a semi-supervised learning model for tagging various elements in an author's profile such as affiliation, phone number etc. Our results on extracting research interests is a first step in this direction.

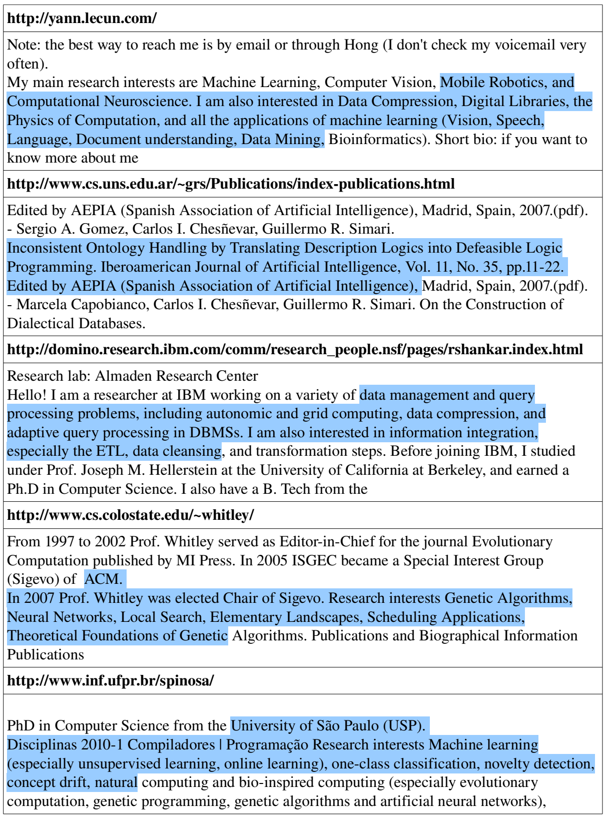
7. REFERENCES

[1] K. Balog and M. De Rijke. Determining expert profiles (with an application to expert finding). In *IJCAI*, 2007.

[2] Z. Bar-yossef and M. Gurevich. Random sampling from a search engines index. In *WWW*, 2006.

[3] L. Becchetti, C. Castillo, D. Donato, and A. Fazzone. A comparison of sampling techniques for web characterization. In *LinkKDD*, 2006.

Figure 5: Example Segments Extracted



http://yann.lecun.com/

Note: the best way to reach me is by email or through Hong (I don't check my voicemail very often).
My main research interests are Machine Learning, Computer Vision, Mobile Robotics, and Computational Neuroscience. I am also interested in Data Compression, Digital Libraries, the Physics of Computation, and all the applications of machine learning (Vision, Speech, Language, Document understanding, Data Mining, Bioinformatics). Short bio: if you want to know more about me

http://www.cs.uns.edu.ar/~grs/Publications/index-publications.html

Edited by AEPIA (Spanish Association of Artificial Intelligence), Madrid, Spain, 2007.(pdf). - Sergio A. Gomez, Carlos I. Chesñevar, Guillermo R. Simari.
Inconsistent Ontology Handling by Translating Description Logics into Defeasible Logic Programming. Iberoamerican Journal of Artificial Intelligence, Vol. 11, No. 35, pp.11-22. Edited by AEPIA (Spanish Association of Artificial Intelligence), Madrid, Spain, 2007.(pdf). - Marcela Capobianco, Carlos I. Chesñevar, Guillermo R. Simari. On the Construction of Dialectical Databases.

http://domino.research.ibm.com/comm/research_people.nsf/pages/rshankar.index.html

Research lab: Almaden Research Center
Hello! I am a researcher at IBM working on a variety of data management and query processing problems, including autonomic and grid computing, data compression, and adaptive query processing in DBMSs. I am also interested in information integration, especially the ETL, data cleansing, and transformation steps. Before joining IBM, I studied under Prof. Joseph M. Hellerstein at the University of California at Berkeley, and earned a Ph.D in Computer Science. I also have a B. Tech from the

http://www.cs.colostate.edu/~whitley/

From 1997 to 2002 Prof. Whitley served as Editor-in-Chief for the journal Evolutionary Computation published by MI Press. In 2005 ISGEC became a Special Interest Group (Sigevo) of ACM.
In 2007 Prof. Whitley was elected Chair of Sigevo. Research interests Genetic Algorithms, Neural Networks, Local Search, Elementary Landscapes, Scheduling Applications, Theoretical Foundations of Genetic Algorithms. Publications and Biographical Information Publications

http://www.inf.ufpr.br/spinosa/

PhD in Computer Science from the University of São Paulo (USP).
Disciplinas 2010-1 Compiladores | Programação Research interests Machine learning (especially unsupervised learning, online learning), one-class classification, novelty detection, concept drift, natural computing and bio-inspired computing (especially evolutionary computation, genetic programming, genetic algorithms and artificial neural networks),

[4] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. In *WWW*, 1998.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 2003.

[6] C. Boulis and M. Ostendorf. Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams. In *FSDM*, 2005.

[7] A. Broder. A taxonomy of web search. *SIGIR Forum*, 2002.

[8] A. Broder, M. Fontura, V. Josifovski, R. Kumar, R. Motwani, S. Nabar, R. Panigrahy, A. Tomkins, and Y. Xu. Estimating corpus size via queries. In *CIKM*, 2006.

[9] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001.

[10] H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on dblp bibliography data. In *ICDM*, 2008.

[11] K. Deng and A. W. Moore. On the greediness of feature selection algorithms, 1998.

[12] A. Dobra and S. E. Fienberg. How large is the world wide web? In *Web Dynamics*. 2004.

[13] E. I. George and C. P. Robert. Capture-recapture estimation via gibbs sampling. In *Biometrika*, 1992.

[14] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 2004.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. In *SIGKDD Explorations, Volume 11, Issue 1*, 2009.

[16] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *JCDL*, 2003.

[17] G. Heinrich. Parameter estimation for text analysis. Technical report, 2008.

[18] J. Huang, S. Ertekin, and C. Giles. Efficient name disambiguation for large-scale databases. In *PKDD*. 2006.

[19] I. Kanaris and E. Stamatatos. Learning to recognize webpage genres. *Inf. Process. Manage.*, 2009.

[20] R. King and S. P. Brooks. On the bayesian analysis of population size. 2001.

[21] S. Lawrence and C. L. Giles. Searching the world wide web. In *Science*. 1998.

[22] H. Li, I. G. Councill, L. Bolelli, D. Zhou, Y. Song, W.-C. Lee, A. Sivasubramaniam, and C. L. Giles. Citeseerx: a scalable autonomous scientific digital library. In *InfoScale '06*, 2006.

[23] S. M. Automatic identification of genre in web pages. *PhD Thesis*, 2007.

[24] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[25] D. M. Mimno and A. McCallum. Mining a digital library for influential authors. In *JCDL*, 2007.

[26] Nadeau, David, Sekine, and Satoshi. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 2007.

[27] R. Nallapati. Discriminative models for information retrieval. In *SIGIR*, 2004.

[28] D. L. Otis, K. P. Burnham, G. C. White, and D. R. Anderson. Statistical inference from capture data on closed animal populations. 1978.

[29] S. Pledger. Unified maximum likelihood estimates for closed capture recapture models using mixtures. 2000.

[30] D. Poole. Estimating the size of the telephone universe: a bayesian mark-recapture approach. In *KDD*, 2004.

[31] X. Qi and B. D. Davison. Web page classification: Features and algorithms. *ACM Comput. Surv.*, 2009.

[32] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 2001.

[33] A. Silvescu, C. Caragea, and V. Honavar. Combining super-structuring and abstraction on sequence classification. In *ICDM*, 2009.

[34] Y. Sun and C. L. Giles. Estimating the web robot population. In *WWW*, 2010.

[35] J. Tang, D. Zhang, and L. Yao. Social network extraction of academic researchers. In *ICDM*, 2007.

[36] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, 2008.

[37] T. Upstill, N. Craswell, and D. Hawking. Query-independent evidence in home page finding. *ACM Trans. Inf. Syst.*, 2003.

[38] Y. Wang and K. Oyama. Web page classification exploiting contents of surrounding pages for building a high-quality homepage collection. In *Digital Libraries: Achievements, Challenges and Opportunities*. 2006.

[39] C. Whitelaw, A. Kehlenbeck, N. Petrovic, and L. Ungar. Web-scale named entity recognition. In *CIKM*, 2008.

[40] W. Xi, E. Fox, R. Tan, and J. Shu. Machine learning approach for homepage finding task. In *String Processing and Information Retrieval*. 2002.

[41] S. Zheng, R. Song, J.-R. Wen, and D. Wu. Joint optimization of wrapper generation and template detection. In *KDD*, 2007.

[42] S. Zheng, D. Zhou, J. Li, and C. L. Giles. Extracting author meta-data from web using visual features. *ICDM Workshops*, 2007.