Disambiguation of People's Web Appearances

My primary interest is automatic identification of web presence of particular people. The problem appears to have a trivial solution when the name of the person is unique. we can just google this person's name. However, the problem complexity grows significantly with the level of commonness of the personal name. Indeed, given a common name such as "Tom Mitchell", we find hundreds of different people called Tom Mitchell. The main complication is that a web mining system cannot a priori know how ambiguous one or another personal name is. Yet another complication is that the person we are looking for may not have *any* web presence, but his or her namesakes have. And even if the person is well presented in the web, an existence of his of her famous namesake can make the search practically impossible (compare "Julia Roberts" and "Julia Roberts" Professor WKU).

Well, given *just* a personal name, the problem of finding the person's web presence cannot be theoretically resolved. Some additional information about the person is required. When we search for a person in the web, we usually construct a query that provides some kind of summary of the person's activities, e.g. "Tom Mitchell" Professor CMU. However, if our query is too common (such as "Tom Mitchell" Professor), our first hit will be Tom Mitchell the UChicago Professor, and if our query is too narrow (as "Tom Mitchell" "Professor CMU"), we may not find anyone. Since modern search engines are so sensitive to small variations of the same query, it is hard to require an automatic system to construct queries of high quality.

We noticed however that the problem of personal name disambiguation becomes much easier if we are given *a few* names of people who are known to be related to each other. They can for example be co-authors of scientific publications, share a profession or a hobby. Even if their names are totally ambiguous, when we google two names together, like "Tom Mitchell" "William Cohen", we are almost guaranteed to find the right people. Or accidentally to find nothing, if web presence of at least one of the two people is sparse enough.

I have built a web mining system that disambiguates web appearances of a group of people. I test the system on a list of 12 personal names. These names are taken from headers of email messages from one folder of a CALO participant's email directory. For each name, I retrieved 100 first Google hits and labeled them manually. After removing some empty pages and error statements, the resulting dataset consists of 1085 web pages that refer to 187 different people, from which 420 pages are relevant (they refer to the 12 particular people we are looking for). Some statistics of the dataset can be found below.

Number of relevant pages	Number of namesakes	Number of retrieved pages	Position	Personal name
96	2	97	SRI Manager	*Adam Cheyer*
6	10	88	CMU Professor	*William Cohen*
64	6	81	SRI Engineer	*Steve Hardt*

20	19	92	SRI Manager	*David Israel*
88	2	89	MIT Professor	*Leslie Pack Kaelbling*
11	8	94	SRI Manager	*Bill Mark*
54	16	94	UMass Professor	*Andrew McCallum*
15	37	92	CMU Professor	*Tom Mitchell*
1	13	94	Stanford Undergrad	*David Mulford*
32	29	87	Stanford Professor	*Andrew Ng*
32	19	88	UPenn Professor	*Fernando Pereira*
1	26	89	SRI Engineer	*Lynn Voss*

The dataset is ready for download. Download the preprocessed dataset (4.8 Mb tarred, gzipped). Makeup language was removed from all the pages. At the end of each page there is a list of its hyperlinks, starting with the URL of the page.

Publication. We proposed two unsupervised methods for the web appearance disambiguation. One is based on link structure of the pages, another one on distributional clustering of the pages. The hybrid of the two methods achieves above 80% F-measure on our dataset. More information can be found in:

Disambiguating Web Appearances of People in a Social Network. Joint work with A. McCallum. *In Proceedings of WWW 2005* pdf .

s