UW People Search

Vidhya Murali

University of Wisconsin Madison vidhyam@cs.wisc.edu

Abstract

This project aims at developing a tool for searching people in a specific domain using text search queries. In addition to producing accurate search results the project also intends to provide a graphical 2-D representation of the people in the domain. Also the search queries are projected in the people clustering plot to better understand the relation between the search results.

Introduction

In today's internet world, search engines have become an integral part of our daily activities. With a dozen and odd general search engines available, searching for information based on keywords/topics has become a simple activity. Most of these search engines use keyword based paradigm for performing search and hence the results turn out to be more in terms of query text occurrence in the document. So currently for users to perform people search, they would need to know the person's name and/or other structured details like address/phone number to look up on the internet. Motivated by the idea of enabling people on the internet to be looked upon based on keywords/tags, this project aims at exploring options for the same.

UW People Search

The People Search is a good combination of an expert finder and a general search engine. Given a topic/tag we expect to obtain a list of most relevant people to the search term. Considering the dataset belongs to a specific domain, we crawl, index and rank only information pages pertaining to the people belonging to this domain. In our case we have chosen the domain of cs.wisc.edu (i.e.) the computer science department within the University of Wisconsin Madison.

There are two parts to producing the results of a query search: (i) Cosine similarity based search result (ii) PCA clustering plot. These have been explained in the following section.

Textual Search Result

This is similar to a regular search result. It is performed by calculating the cosine similarity between the query term and the people's feature vector in the domain. The technique of TFIDF has been employed for calculating the term weights for each individual person in the domain.

In addition to cosine similarity score, words' co-occurrence weights and window distance between the query terms in the person's documents have been used to determine the ranks of the search results

PCA based plot

The second part of the people search result is a visual representation of the search results. The input query is projected in a 2 dimensional plot. The technique of PCA (principal component analysis) has been applied to perform dimension reduction of the original feature vectors. A plot representing each individual person in the domain is plotted using the two principal components obtained from PCA. The given search query is then projected on this plot. This graphical representation helps to understand the distribution of people around the given search term as well as compare and contrast query results.

Input Corpus

Unstructured data of peoples' web pages in cs.wisc.edu domain forms the input corpus. These pages are crawled and categorized into Faculty, Staff, Grad students and Undergrad students. This categorization is based on the base url of the people. (Eg: faculty list is extracted from the www.cs.wisc.edu/faculty.html)

In all there are 234 people in the domain crawled and 8371 word types in the vocabulary obtained after performing the following text PREPROCESSING steps.

- Html stripping
- Unigram model(bag of words)
- Stop word filtering
- Proper noun identification
- Case folding

A matrix of people against vocabulary's word types is also computed for performing PCA in Matlab.

Results

The top results of some search query is tabulated below with their corresponding meta details (scores and tags associated with the search result)

Search	Faculty Search	Staff Search Result	Grad student Search	UnderGrad student
term	Result		result	Search Result
Artificial intelligen ce	Xiaqin (Jerry) Zhu [scored: 802902902 963, teg* intelligence erificial] Jude W. Shavili [score=0: 302150537 634, teg* intelligence artificial]		Shenoy, Pete Priyananda [score-0.907843137 285, tag- intelligence artificial] Par, Yue [score-0.807207207 207, tag- intelligence artificial]	Lenz, Mark [s.core=0.804278074 988, tag= intelligence artificial] Ruberstein, Abraham [s.core=0.803238866 387, tag= intelligence artificial]
Jerry	Xisojin (Jerry) Zhu [score=1, teg=jerry]		Fillmore, Nathanael [score=0, 601425641 026, tag= jerry] Eldawy, Mohamad [score=0, 601380392 157, tag= jerry]	
Machine	Jude W. Stevik	_	Berg Elizabeth	Byrne, John
Learning	[score=0.798201147 477. bg= machine learning] AnHai Doan [score=0.790259025 09. tag= machine learning] Xiagin (Jerry) Zhu [score=0.778812479 07, tag= machine learning]		[score=0.808510838 230, seg-machine learning] Sun,Chong [score=0.807547108 811, tag-machine learning] Fan,Yue [score=0.807207207 207, seg-machine learning]	[score=0,569568963 188; tag= machine leaming] Ruberstein, Abraham [score=0,425473785 311, tag= machine]
Condor	Miron Livny [score=0.604701075 269, teg= condor]	Werner Virginia [score=0.000909572 238, tag= condor]	Alderman, lan [score=0, 607672727 273, tag= condor] Paulson, Erik [score=0, 601498901 098, tag= condor]	
Wireless Networks	Suman Banerjee [score=0.616979742 173, tag= wireless networks] Aditya Akella [score=0.647261943 118, tag= wireless networks]		Ramonurthy,Pratap [score=0.803624229 075, tage wireless networks] Sen,Sayandesp [score=0.803187250 990, tage wireless networks]	Projnek Mary [score=0.425891062 577, tag= networks]
Coordinat		Thorp Angels [score=0.600884261 501, tag= coordinator]		

Table 1: Table of search results

A PCA 2-D representation of the domain people is given below.

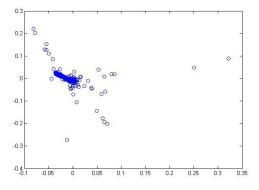


Figure 1: 2-d plot of people in the domain

This plot is obtained by considering the Eigen vector of covariance matrix corresponding to maximum variance (high Eigen values). The original feature vector of the people in the domain is reduced to the top two Eigen vector dimensions and represented on a 2-D plot.

Given below is the plot with the projection of search queries in Table 1 included.

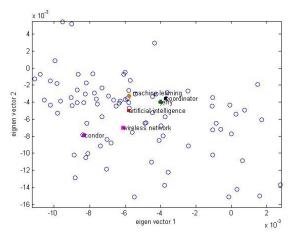


Figure 2: zoomed view of 2-d plot with queries in Table 1 projected

CONCLUSION

While the results produced are satisfactory, they have not been compared against any existing similar tools or standards to determine precision of this system.

Also a better idea about this systems performance and accuracy of results can be obtained by extending this tool to a bigger dataset (domain).

Limitations

Considering the dataset being unstructured data (web pages) the search results depends on the extraction technique used on the data sources to obtain the features.

Also automation of identifying and crawling new datasources within the domain is another interesting challenge.

The graphical representation produced has some loss of data. When the multidimensional feature vector is reduced to two dimensions, finer details are lost. So the graphical representation should be viewed as good approximation of the people clustering patterns seen in the domain.

Applications

Domain based people search has far and out reaching application. With online social networks, blogs and other online user interactive groups getting popular, a search tool that can perform search on people based on topics and keywords as apposed to the traditional people search based on first and last name would be very helpful.It can help see commonalities between people and other interesting patterns.

References

- 1. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C., & Nevill-Manning, C.G. (1999). Domain-specific key phrase extraction. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999, 668-673.
- 2. Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2003, 216-223.
- 3. Jonathon Shlens, Center for Neural Science, New York University and Systems Neurobiology Laboratory, Salk Institute for Biological Studies, 2009, A Tutorial on Principal Component Analysis
- 4. Matlab documentation for Principal component analysis www.mathworks.com/access/helpdesk/help/toolb ox/stats/**princomp**.html