University of California, San Diego

iDASH Internship Report

Trends in biomedical informatics: Topic analysis of JAMIA articles

Author:
Chao Jiang

Supervisor:
Dr. Xiaoqian Jiang/ Dr. Shuang Wang

Monday 21st September, 2015

1 Introduction

The aim of this project is to investigate the topic trend of articles related with biomedical informatics. It covers the methods used to collect data from websites and analyses some results.

2 Methods

2.1 Collecting Data Technique

We retrieved Journal of the American Medical Informatics Association(JAMIA) articles published between 2009 and 2014 from the Institute for Scientific Information's (ISI) Web of Science database[1]. We included a total of 941 articles in this study. For each paper we implement the python script[2] to crawl data from different websites. So we can get the title, authors, published date, citation for each year, reference papers etc. of each article .

2.2 Analyzing Data

In this study, we built two three-way tensors to study the trends between topics and citations using tensor factorization. The first dimension of both tensors is time. The second dimension of both tensors presents citations per month (CPM). The third dimension of the first tensor included the nine most frequent topic categories as we had utilized in previous work. another tensor's third dimension represented 1,417 MeSH terms from 941 articles, which can be used to extract latent factors without restriction to nine pre-selected categories. Figure 1 conceptually illustrates an example of a three dimensional tensor and its decomposition process using tensor factorization.

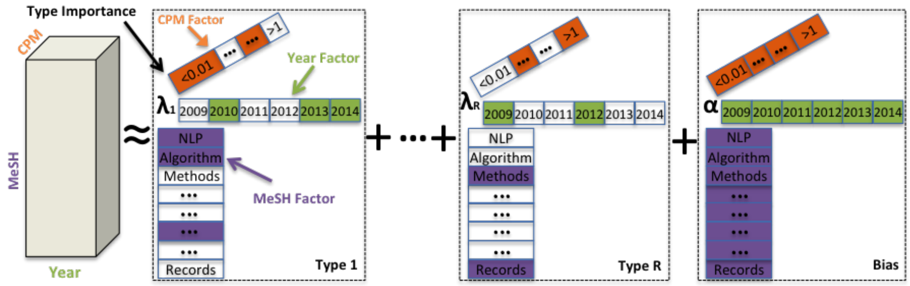


Figure 1: An example of a three dimensional tensor and its decomposition process using tensor factorization

[1] http://webofscience.com

[2] https://github.com/chao92/webcraw_python

3 Results

Two experiments had been done for tensor factorization. Before we done tensor factorization, we plot a figure to showed the breakdown of the Percentage of Articles Published in each Year (PAPY) in line plots (right vertical axis), as well as the Average number of Citations Per Article (ACPA) in each year in bar plots (left vertical axis). As figure 2 below. From this figure, we can get that Electronic Health Records always lead the popularity and the number of articles in Methods and Algorithms are also increasing dramatically within 2013 and 2014. Next we will prove our result above by using tensor factorization. Figure

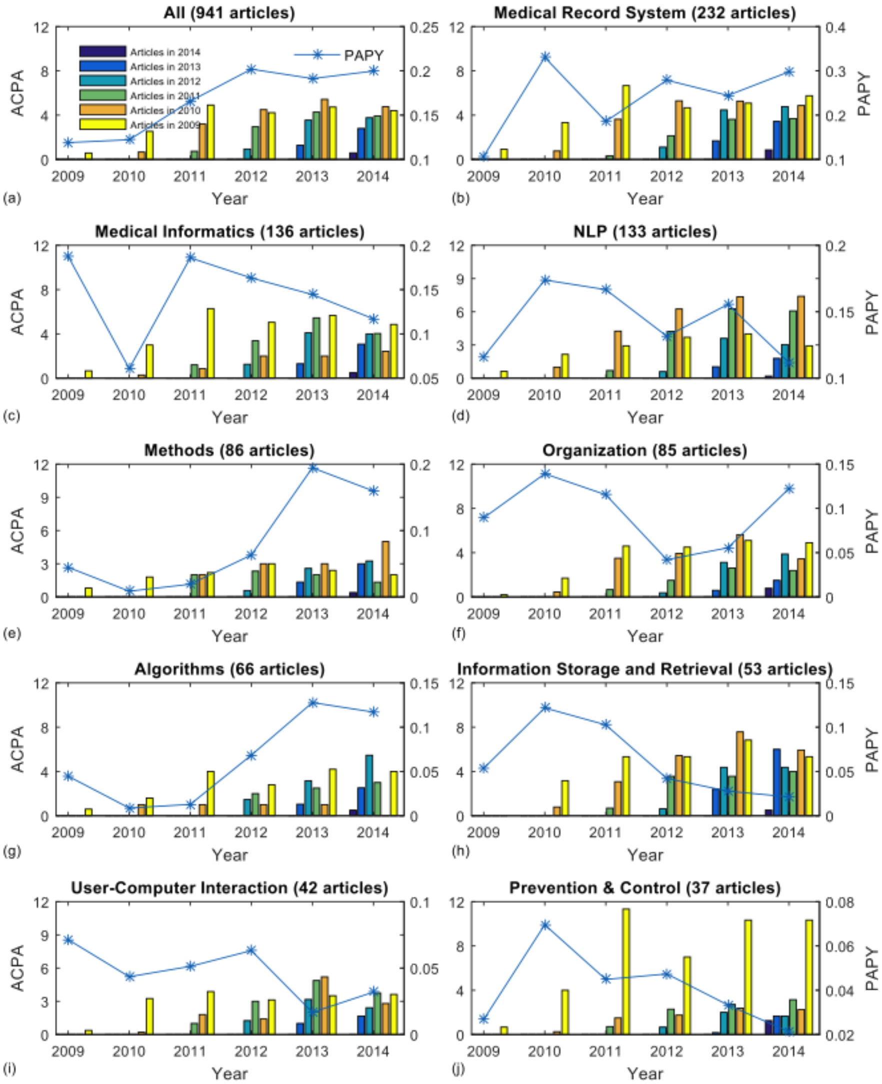
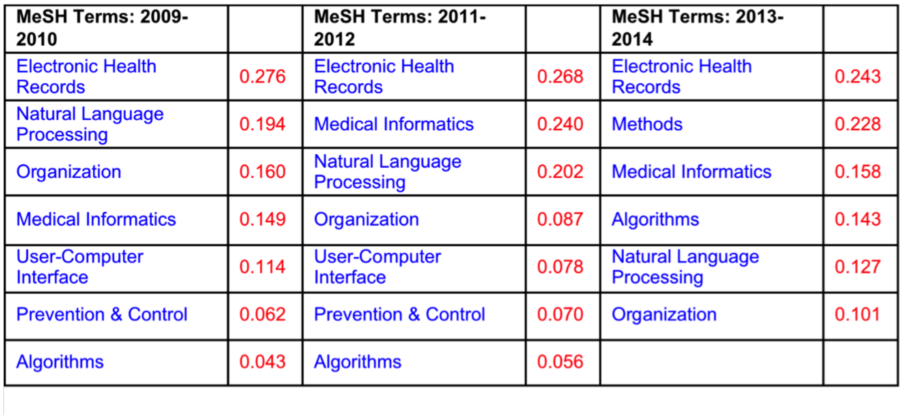


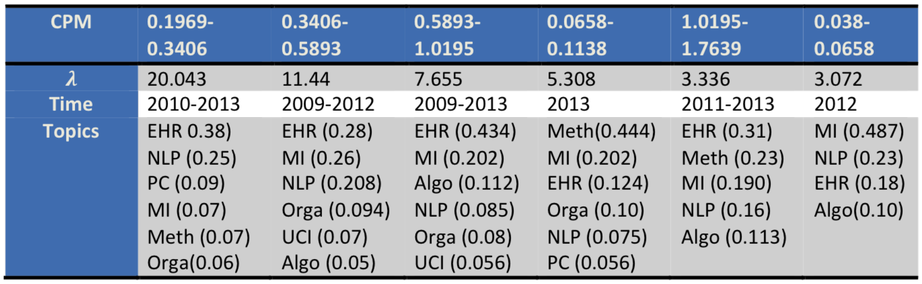
Figure 2: Tends of nine categories articles in JAMIA

3 is Decomposed tensors using year-guided tensor factorization,where categories in a decomposed tensor are listed in descending order of weight, representing frequently appearing topics. The next following two figures 4 and 5 are the result of tensor factorization (TF) based on 9 categories. from these two figures we can know that the most paper's CPM values located between 0.113 and 1.11 at the same time most of them are related with Electronic Health Records. Also papers concerned with NLP and published in the second half of 2010 have the highest value of CPM.



MeSH Terms: 2009-2010		MeSH Terms: 2011-2012		MeSH Terms: 2013-2014	
Electronic Health Records	0.276	Electronic Health Records	0.268	Electronic Health Records	0.243
Natural Language Processing	0.194	Medical Informatics	0.240	Methods	0.228
Organization	0.160	Natural Language Processing	0.202	Medical Informatics	0.158
Medical Informatics	0.149	Organization	0.087	Algorithms	0.143
User-Computer Interface	0.114	User-Computer Interface	0.078	Natural Language Processing	0.127
Prevention & Control	0.062	Prevention & Control	0.070	Organization	0.101
Algorithms	0.043	Algorithms	0.056		

Figure 3: Interaction tensors after applying tensor factorization on an 11(CPM) × 12(half year) × 9 (categories) tensors



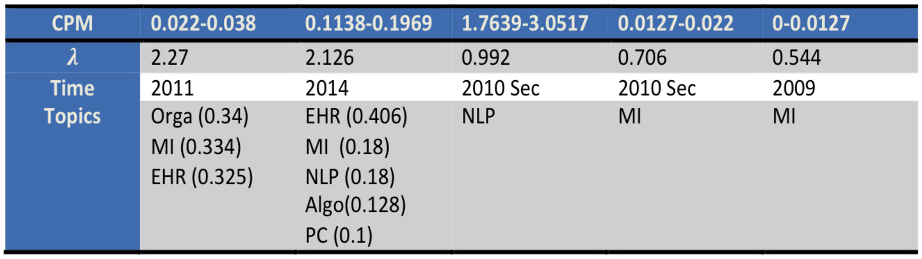
CPM	0.1969-0.3406	0.3406-0.5893	0.5893-1.0195	0.0658-0.1138	1.0195-1.7639	0.038-0.0658
λ	20.043	11.44	7.655	5.308	3.336	3.072
Time	2010-2013	2009-2012	2009-2013	2013	2011-2013	2012
Topics	EHR 0.38) NLP (0.25) PC (0.09) MI (0.07) Meth (0.07) Orga(0.06)	EHR (0.28) MI (0.26) NLP (0.208) Orga (0.094) UCI (0.07) Algo (0.05)	EHR (0.434) MI (0.202) Algo (0.112) NLP (0.085) Orga (0.08) UCI (0.056)	Meth(0.444) MI (0.202) EHR (0.124) Orga (0.10) NLP (0.075) PC (0.056)	EHR (0.31) Meth (0.23) MI (0.190) NLP (0.16) Algo (0.113)	MI (0.487) NLP (0.23) EHR (0.18) Algo(0.10)

Figure 4: First six interaction tensors after applying TF

4 Conclusion

In this summer, we developed Python scripts to automatically retrieve JAMIA publication data. and through applying tensor factorization, we are able to get more detail information then before. We have submitted our paper[3] to JAMIA(2015)

[3] Dong Han, Shuang Wang, Chao Jiang, Xiaoqian Jiang, HyeonEui Kim, Jimeng Sun, Lucila Ohno-Machado "Trend in biomedical informatics: Topic analysis of JAMIA articles" (JAMIA)



CPM	0.022-0.038	0.1138-0.1969	1.7639-3.0517	0.0127-0.022	0-0.0127
λ	2.27	2.126	0.992	0.706	0.544
Time	2011	2014	2010 Sec	2010 Sec	2009
Topics	Orga (0.34) MI (0.334) EHR (0.325)	EHR (0.406) MI (0.18) NLP (0.18) Algo(0.128) PC (0.1)	NLP	MI	MI

Figure 5: The five interaction tensors after applying TF

5 Future Projects

In our future, we can considering more factors such as authors, institutions,etc. also by applying tensor factorization to patient data to dig more information.