RESEARCHER PROFILING: FINDING REPRESENTATIVE PHRASES FOR RESEARCHERS

By

Kangyan Zhou

Senior Thesis in Computer Engineering

University of Illinois at Urbana-Champaign

Advisor: Kevin Chang

December 2016

Abstract

We are working on building a comprehensive search system for a researcher given his/her name and affiliation. The output result includes the researcher's basic profile, his/her research publications, past grants received, patents, and Youtube or any other video links. In this paper, we utilize an existing framework and propose a method to accurately generate meaningful and representative phrases for one researcher, based on his/her publication titles from the search results of the aforementioned system. The purpose of the research is to provide a thorough understanding of the researcher's interest based on limited input. Although the algorithm requires some background context given the limited size of input, the quality of the phrases generated is satisfactory. We also discuss our approach to generate personalized phrase representation for two or more researchers working in a similar field.

Subject Keywords: key phrase generation; research focus analysis

Acknowledgments

Thanks to my advisor Professor Kevin Chang and my teammates Hai Hu, Hong Cheng, Zhao Weng, and Zubin Pahuja.

Contents

1. Introduction

Generating meaningful and representative phrases for a researcher are not an available feature for popular researcher search systems such as Google Scholar and dblp, though they have various kinds of usage in application context. The phrases can be used for users of our Researcher Profiling system to quickly recognize the major focus of the query researcher. Otherwise even a knowledgeable user may have to read through all the publication titles gathered in the system to get a brief overview and generalization. The phrases can also be used as an important feature to create a good recommendation system for any fixed query researcher by finding researchers with similar key phrases as a major focus.

The task faces several challenges. The inputs are only a number of titles from researcher's publications. For famous researchers, the number can be as large as 100-200, a feasible quantity to use frequent pattern mining based techniques to investigate into the titles and find the meaningful phrases. For average or young researchers, however, the number of titles is limited to 20-50, so any available phrase mining technique would suffer from lack of input data. The researcher may have several major focus areas: finding out representative phrases for all of the areas would be a hard problem. Finally, a researcher's publications may include some phrases that are meaningful in a global context, but not very related to that researcher's focus. For example, Jiawei Han is a famous professor in the data mining field. In his publications, "face recognition", a meaningful word in computer vision field, appears only once. In this case, the phrase should not be picked as a categorized phrase.

In this work, we propose a method to find representative phrases that can categorize a researcher's publications given the researcher's subject, such as Computer Science and Chemistry. Our method can locate most of meaningful phrases that are related to researcher's major area, given only the researcher publication titles. Although the method requires some background data collection and preprocessing, we believe the cost is acceptable.

We also spend some time exploring approaches to generate detailed representative phrases given two researchers working in a similar field. Although the algorithm can generate some meaningful results, it has trouble defining the depth of each topic, and we do not have enough time to improve our approach. We would like to document our method for future research.

2. Literature Review

There are already many related work about phrase segmentation. To finding the representative phrases for one researcher, we find [1], and its implementation, SegPhrase, particularly useful in our context. The input for the system is a collection of short texts, and the output is the meaningful phrases generated from the input. An example of the output is shown in Table 1. The system requires very few or even no predefined labels as extra input. Each output phrase is associated with a score ranging from 0-1, indicating its importance.

Table 1 Example of SegPhrase Output

Phrase	Score
Bayesian network	0.856624
Discriminant Analysis	0.856467
Big Data	0.856132
Parallel Algorithm	0.8534662
Artificial Neural Network	0.853466

The algorithm consists of three parts. It first runs a common frequent pattern mining algorithm to preprocess the input documents and remove the infrequent patterns. It also builds a classifier based on predefined labels and some information retrieval metrics as features such as pointwise mutual information (PMI), pointwise Kullback-Leibler divergence (PKL) and inverse document frequency (IDF). The key phase in the algorithm is the phrasal segmentation part, where a viterbi training algorithm is designed to find the possible best split point of an input titles and get a “rectified” frequency of each phrase. A “rectified” frequency means a good phrase, like “database management system” should increase its frequency based on the raw frequency it has, while a bad phrase, such as “vector machine”, should have a decreased frequency, and eventually a corresponding good and meaningful phrase “support vector machine” should be favored. There is also a feedback learning part to help better estimate the segmentation feature.

Since the output phrases look perfect and there is already a score for each phrase for ranking, our task is transformed to collect a certain amount of titles, use SegPhrase to build a comprehensive background knowledge base, and try to query the knowledge base whenever the user searches a new researcher.

For generating detailed representative phrases for two researchers in a similar field, we naturally think about recursively splitting the topics into subtopics, and trying to find which subtopic each researcher belongs to. There is also an existing work, [2], and its implementation, CATHY, in the field. See Table 2 as an example of CATHY output. The system takes a set of titles as input, and automatically generates a topical hierarchy and meaningful phrases associated with each subtopic.

Table 2 Example of CATHY Output

Topic	Phrase
1	social networks; web page; search engine; information retrieval
2	neural networks; natural language; knowledge discovery
3	decision tree; text classification; hidden markov model
4	data management; query processing; materialized views; data integration
5	clustering data; data mining; time series; nearest neighbor

Essentially the algorithm can be split into two steps. The first step is the recursive mining part. It first does a preprocessing step to stem the input document and removes the infrequent words, just as the first step in Apriori frequent mining algorithm, models the probability of two words appearing in each topic as Poisson distribution, and utilizes the expectation maximization (EM) algorithm to recursively estimate the co-occurrence probability of two words in each topic. It will assign the two words into topics, and recursively going down until a predefined max depth is reached or the remaining documents are not large enough to be splitted again. In this way the algorithm builds a hierarchical topical tree. The second step is a frequent pattern mining for each node in the topic tree. The algorithm allows a trade-off between a max pattern and a closed pattern, and also introduces some of the features for ranking purposes such as coverage, phraseness, and purity. Combining these three metrics, the algorithm can assign a rank for each output phrase.

With the algorithm in hand, and given a field, we can recursively search, for each of the two or more query researchers, the node he/she belongs in the hierarchical topic tree, and try to figure out a way to decide the corresponding output.

3. Description of Research Result

3.1 Representative Phrases for One Researcher

3.1.1 Method Enhancement

For the Computer Science (CS) field, we choose dblp, a computer science bibliography, as our background knowledge base. We feed all the collected titles into SegPhrase after some preprocessing. The output of SegPhrase is stored on the disk with phrases and their corresponding rankings. When a researcher in CS field is queried, all the bigrams and trigrams from his/her titles are generated, and the first 5 to 10 phrases that appear in both the query researcher titles and the background meaningful phrases and have the highest ranking serves as the output.

This method, however, fails to deal with the case as mentioned before in the introduction, where “face recognition” appears in the meaningful phrases of Jiawei Han, an expert in the data mining field. To resolve this issue, we make the assumption that a researcher can only focus on several related fields, without abruptly changing his/her research interest in a his/her career. Thus it is highly likely that if we classify, or in other word, constrain the researcher to certain number of topics in a larger field, the algorithm can output a better result by only matching the meaningful phrases that appear in the assumed topics.

Following this assumption, we run a Latent Dirichlet Allocation (LDA) topic model on the collected titles. The LDA output contains a matrix holding the probability of the appearance of each unigram in each topic. Using this matrix, and with the assumption that each topic has roughly the same frequency in the collected titles, for each title we can build a Naive Bayes classifier with Maximum Likelihood (ML) to generate a 1*N matrix to denote the likelihood of the topic belonging, where N is the number of predefined topics. Thus we can first split the collected titles into the predefined number of topics by picking the topic with the largest probability in the 1*N matrix, and then run SegPhrase to obtain the meaningful phrases for each category.

In order to deal with the case when each title may have multiple topics, we borrow the idea of bagging in machine learning and try to assign a title to several topics simultaneous, or to a topic multiple times if the title is very likely to belong to this topic. The way to decide how many times each title will be assigned to one topic is purely based on the output of the Naive Bayes classifier. Suppose we have N topics. If in the 1*N matrix, there is an entry larger than k/N, where k is an integer larger than 1, we assign the title to that topic k times. In this way we effectively minimize the side effect that each title may belong to several topics.

In the new schema, when there is a new query researcher, we would first classify the researcher into first three most likely topics. Then for each topic, try to find any matching representative phrases in all of the possible bigrams and trigrams appeared in the input query titles and the meaningful phrases generated by SegPhrase.

3.1.2 Data Collection and Result

To construct the background language model, we collect over 100 thousand titles from dblp, available directly from download. Then we parse the xml file and preprocess the titles by deleting all the punctuation, drop all stop words and changing all alphabets to lower case. To reduce a unigram to its base form, we try to stem the word by identifying all the unigrams that were in plural form, ”-ing” form, and in adverb form. I use an English word dictionary and search through the whole dictionary every time we suspected the word is not in its base form. If there is a match, change the original word to the word found in the dictionary. By doing this the vocabulary size reduces to almost 2/3 of the original size, which was very helpful in our task since SegPhrase uses word2vec, relying on the count of appearances of single word. During the preprocessing step, we also find out that some particular words, such as "computing" and "mining", should not be reduced to their base form, since they had very special meanings in computer science area. So we read through the high-frequency words and manually create a list of words with instances of such kind. The number of topics is fixed to 15.

For test set we manually crawl the google scholar webpage, and collect titles about 15 researchers(see Table 3 for the researchers' names and their field). The researchers have various background and cover almost every popular aspect for computer science, such as data mining, machine learning, information retrieval, database systems, natural language processing, compiler and formal method, distributed system, artificial intelligence, computer vision and bioinformatics and computational biology. Some of the researchers are well-known and they have a lot of publications, and we collected about 50 titles for them. Other researchers are still in early stages and do not have many publications, so I collected about 10-20 titles, almost all of their collections. Although the test set had limited size, the coverage was comprehensive and thus remained its informativeness.

Table 3 Examples of Computer Science Researchers and Their Field

Name	Field
Jiawei Han	Data Mining
Jialu Liu	Data Mining
David Forsyth	Computer Vision
Li Fei-Fei	Computer Vision
Chengxiang Zhai	Information Retrieval
Dan Roth	Machine Learning; Natural Language Processing
Gul Agha	Programming Language; Formal Methods; Software Engineering
Svetlana Lazebnik	Computer Vision
Indranil Gupta	Distributed System
Christopher Manning	Information Retrieval; Natural Language Processing
Jeff Erickson	Theory and Algorithms
Saurabh Sinha	Bioinformatics and Computational Biology
Grigore Rosu	Programming Language; Formal Methods; Software Engineering

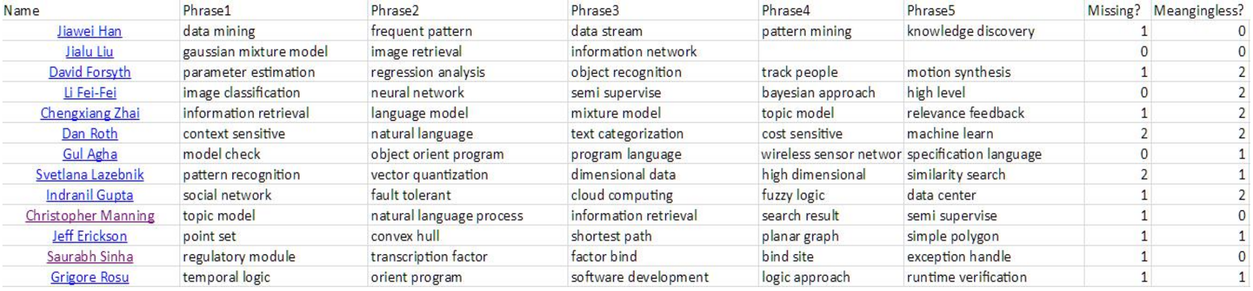
Table 4 Examples of Computer Science Researchers and Their Representative Phrases

Name	Field
Jiawei Han	data mining, frequent pattern, data stream, pattern mining, knowledge discovery
Jialu Liu	Gaussian mixture model, image retrieval, information network
David Forsyth	parameter estimation, regression analysis, object recognition, track people, motion synthesis
Li Fei-Fei	image classification, neural network, semi supervise, bayesian approach, high level
Chengxiang Zhai	information retrieval, language model, mixture model, topic model, relevance feedback
Dan Roth	context sensitive, natural language, text categorization, cost sensitive, machine learn
Gul Agha	model check, object orient program, program language, wireless sensor network, specification language
Svetlana Lazebnik	pattern recognition, vector quantization, dimensional data, high dimensional, similarity search
Indranil Gupta	social network, fault tolerant, cloud computing, fuzzy logic, data center
Christopher Manning	topic model, natural language process, information retrieval, search result, semi supervise
Jeff Erickson	point set, convex hull, shortest path, planar graph, simple polygon
Saurabh Sinha	regulatory module, transcription factor, factor bind, bind site, exception handle
Grigore Rosu	temporal logic, orient program, software development, logic approach, runtime verification

From the result shown in Table 4, we can see that the representative phrases for most of the researchers are related to their major field, acquired from area of interest of the corresponding university websites.

There are some difficulties in evaluating the results. The common metrics used is precision, defined as True Positive (TP)/(True Positive (TP) + False Positive (FP)), and recall, defined as True Positive (TP)/(True Positive (TP) + False Negative (FN)). However, in this task, both precision and recall are hard to define. Each evaluator may have their own knowledge or perception about the correctness of the phrases and focus of the researchers, so in general the evaluation might be hard to explain.

We ask students studying Computer Science/Computer Engineering as domain experts to evaluate our result. We design an excel form as shown in figure 1 to ask the students to fill in the “Missing” and “Meaningless” column, where “Missing” is used to evaluate recall and “Meaningless” is used to evaluate precision. The result are shown in Table 5.



Name	Phrase1	Phrase2	Phrase3	Phrase4	Phrase5	Missing?	Meangingless?
Jiawei Han	data mining	frequent pattern	data stream	pattern mining	knowledge discovery	1	0
Jialu Liu	gaussian mixture model	image retrieval	information network			0	0
David Forsyth	parameter estimation	regression analysis	object recognition	track people	motion synthesis	1	2
Li Fei-Fei	image classification	neural network	semi supervise	bayesian approach	high level	0	2
Chengxiang Zhai	information retrieval	language model	mixture model	topic model	relevance feedback	1	2
Dan Roth	context sensitive	natural language	text categorization	cost sensitive	machine learn	2	2
Gul Agha	model check	object orient program	program language	wireless sensor networ	specification language	0	1
Svetlana Lazebnik	pattern recognition	vector quantization	dimensional data	high dimensional	similarity search	2	1
Indranil Gupta	social network	fault tolerant	cloud computing	fuzzy logic	data center	1	2
Christopher Manning	topic model	natural language process	information retrieval	search result	semi supervise	1	0
Jeff Erickson	point set	convex hull	shortest path	planar graph	simple polygon	1	1
Saurabh Sinha	regulatory module	transcription factor	factor bind	bind site	exception handle	1	0
Grigore Rosu	temporal logic	orient program	software development	logic approach	runtime verification	1	1

Figure 1 Screenshot of Form Used for Evaluation

Table 5 Examples of Computer Science Researchers and Their Precision and Recall

Name	Precision(Round to nearest 0.1)	Recall(Round to nearest 0.1)
Jiawei Han	0.8	0.6
Jialu Liu	1	1
David Forsyth	1	0.6
Li Fei-Fei	0.8	0.5
Chengxiang Zhai	1	0.6
Dan Roth	0.8	0.5
Gul Agha	0.8	0.5
Svetlana Lazebnik	0.6	0.4
Indranil Gupta	0.8	0.5
Christopher Manning	1	0.8
Jeff Erickson	1	0.8
Saurabh Sinha	0.8	1
Grigore Rosu	1	1

The precision is very high, which means most of the phrases we output can correctly identify the interest of the given researchers. Some of the recalls are very high, but some of the recalls are very low. In general, researchers working in popular fields, such as computer vision and data mining, with which students are more familiar, have a lower recall. On the other hand, the students are not very familiar with the researchers focusing on other domains, so the recalls are somewhat higher.

The method fails to generate enough meaningful information for cross-disciplinary researchers. As an example, Professor Saurabh Sinha, an expert in Bioinformatics and Computational Biology. The collected titles, however, only focus on Computer Science field. Therefore they may not contain enough information to extract meaningful phrases for the Biology field. And students do not have enough background knowledge to identify this issue.

We also try to extend our work to the Chemistry field. However, since there is no such comprehensive bibliography in Chemistry field. So we manually collect the publication titles on Google scholar of the top 10 universities based on US News ranking. The size of title is about 15000. Table 6 shows the name and field for researchers we use for testing. Table 7 shows the result of query.

Table 6 Examples of Chemistry Researchers and Their Field

Name	Field
Paul V. Braun	Nano and Microstructures
Martin D. Burke	Organic Chemistry
Jefferson Chan	Protein Molecule
Scott E. Denmark	Organic Synthesis
Dana D. Dlott	Energy Storage
Alison R. Fout	Catalysis
Robert B. Gennis	Biochemistry
John A. Gerlt	Organic Chemistry
Andrew A. Gewirth	Electrochemical Study
Gregory S. Girolami	Organic Synthesis
Steve Granick	Polymer Science
Martin Gruebele	Protein Folding
Sharon Hammes-Schiffer	Enzyme
Paul J. Hergenrother	Drug Delivery
So Hirata	Vibrational Energy

Table 7 Examples of Chemistry Researchers and Their Representative Phrases

Name	Phrases
Paul V. Braun	solar cell, photonic crystal, three dimensional, inverse opal, thermal transport
Martin D. Burke	
Jefferson Chan	transcription factor, gene expression, membrane protein, small molecule
Scott E. Denmark	lewis acid, palladium catalyze, aryl halide, organic synthesis
Dana D. Dlott	vibrational energy, energy transfer, raman spectroscopy, carbon dioxide, vibrational spectroscopy
Alison R. Fout	crystal structure, olefin metathesis, active site, lewis acid, aryl halide
Robert B. Gennis	cytochrome oxidase, escherichia coli, active site, electron transfer, transfer reaction
John A. Gerlt	escherichia coli, active site, crystal structure, amino acid, hydrogen bond
Andrew A. Gewirth	electronic structure, scan tunnel microscopy, vibrational spectroscopy, electron transfer, excite state
Gregory S. Girolami	transition metal, nuclear magnetic resonance, excite state, crystal structure
Steve Granick	block copolymer, hydrogen bond, self assembly, liquid film, confine liquid
Martin Gruebele	protein fold, laser spectroscopy, transition state, free energy, absorption spectroscopy
Sharon Hammes-Schiffer	couple electron transfer, electron transfer reaction, potential energy, electron transfer
Paul J. Hergenrother	small molecule, estrogen receptor, natural product
So Hirata	density functional theory, growth factor, nitric oxide, digital sky survey

We also apply the same evaluation technique by asking university students in Chemistry major as domain experts to evaluate our result.

Table 8 Examples of Chemistry Researchers and Their Precision and Recall

Name	Precision(Round to nearest 0.1)	Recall(Round to nearest 0.1)
Paul V. Braun	0.6	0.8
Martin D. Burke	0.8	1
Jefferson Chan	0.8	0.8
Scott E. Denmark	0.8	1
Dana D. Dlott	0.8	1
Alison R. Fout	0.6	0.6
Robert B. Gennis	0.6	0.6
John A. Gerlt	1	1
Andrew A. Gewirth	1	1
Gregory S. Girolami	0.8	1
Steve Granick	0.8	1
Martin Gruebele	0.8	1
Sharon Hammes-Schiffer	0.8	1
Paul J. Hergenrother	0.8	0.8
So Hirata	0.8	1

From the result, we can see that the precision for the researchers, in average, is lower than tat of Computer Science researchers. The recall, however, is much higher. The result might be related to the number of publications of researchers. In Computer Science section, most of the researchers are famous and they have been doing research for a very long time, thus they many different research interests, so the precision would be intrinsically higher, since there are many phrases that can represent their research work; the recall, however, would be lower since most of the time five representative phrases are not enough to generalize the researcher’s work. The Chemistry researchers we use for testing have somewhat opposite condition, so we believe the result are explanatory.

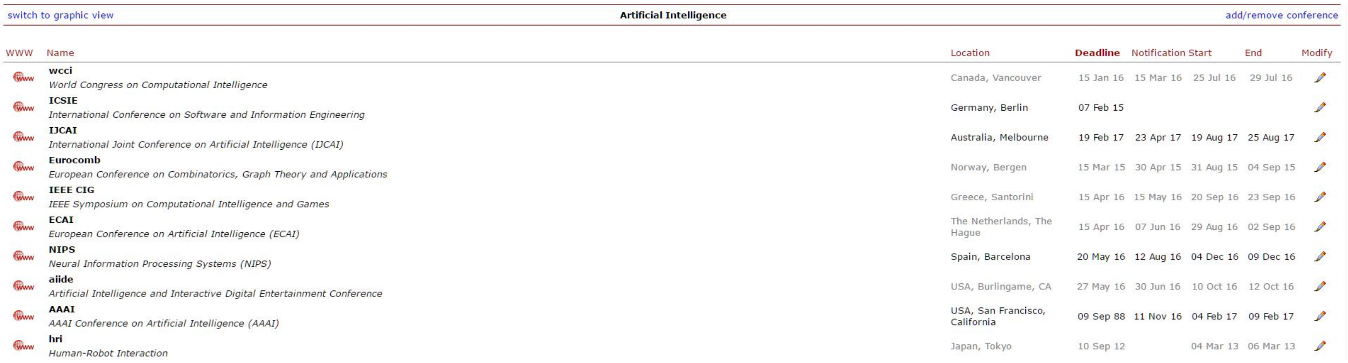
3.2 Representative Phrases for Two Researchers

3.2.1 Method Enhancement

To generate representative phrases for two researchers in the Computer Science field, we first feed the titles we collected in the first part into the CATHY Framework as suggested. When we have two query researchers, we try to find the lowest common ancestor of the researchers first, and then output the meaningful phrases of the child node of the ancestor, where each researcher belongs. The result, however, is not satisfactory, because using the two-word occurrence linkage as the basic unit in an EM algorithm might lead to domain overlapping, thus generating some unreasonable results. For example, the word "network" is more likely to appear in the communication network field. However, with the development of neural networks, phrases like convolutional neural network and recurrent neural network have become popular in the fields of computer vision or natural language processing field. But when we simply feed all the titles in, one topic dominated by work 'network' would be mixed with phrases in both communication network field and computer vision field. Thus the hierarchy is too hybrid to get enough meaningful results.

To resolve this issue, we decide to use cleaner input for the system. For this purpose we utilize the titles of Computer Science conferences for each predefined category. We find a conference search website, [3], that matches out purpose. The website, as the Figure 2 shows, categorizes the most famous Computer Science conferences into different topics. After we identify the conferences for each topic, we utilize the dblp API, as figure 3 shows, to collect the titles of the conference. After collecting the all the titles for the conferences for each category, we feed the titles in each category into the CATHY framework, and generate hierarchy for each topic.

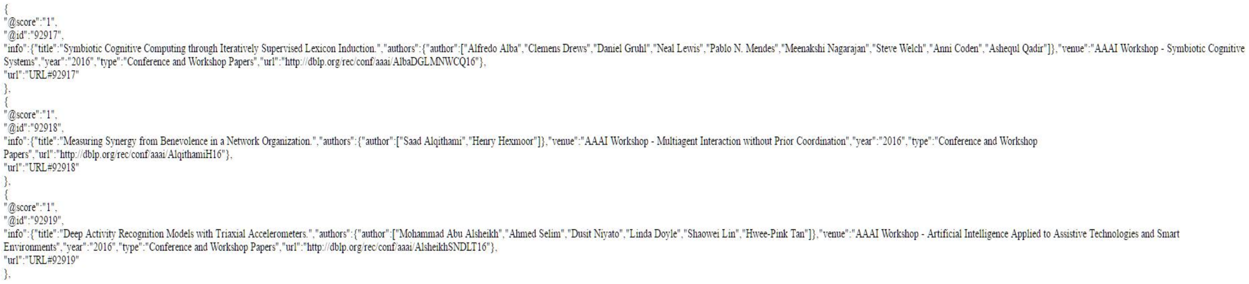
To classify the query titles into predefined category, we build a Naive Bayes classifier for our purpose. The idea is very similar to the naive bayes classifier built in section 3.1.1, except that we use Maximum A Posteriori (MAP) to produce the final estimation. The prior probability is shown in Table 8.



switch to graphic view | **Artificial Intelligence** | add/remove conference

WWW	Name	Location	**Deadline**	Notification	Start	End	Modify
www	**wcci** *World Congress on Computational Intelligence*	Canada, Vancouver	15 Jan 16	15 Mar 16	25 Jul 16	29 Jul 16	✎
www	**ICSIE** *International Conference on Software and Information Engineering*	Germany, Berlin	07 Feb 15				✎
www	**IJCAI** *International Joint Conference on Artificial Intelligence (IJCAI)*	Australia, Melbourne	19 Feb 17	23 Apr 17	19 Aug 17	25 Aug 17	✎
www	**Eurocomb** *European Conference on Combinatorics, Graph Theory and Applications*	Norway, Bergen	15 Mar 15	30 Apr 15	31 Aug 15	04 Sep 15	✎
www	**IEEE CIG** *IEEE Symposium on Computational Intelligence and Games*	Greece, Santorini	15 Apr 16	15 May 16	20 Sep 16	23 Sep 16	✎
www	**ECAI** *European Conference on Artificial Intelligence (ECAI)*	The Netherlands, The Hague	15 Apr 16	07 Jun 16	29 Aug 16	02 Sep 16	✎
www	**NIPS** *Neural Information Processing Systems (NIPS)*	Spain, Barcelona	20 May 16	12 Aug 16	04 Dec 16	09 Dec 16	✎
www	**aiide** *Artificial Intelligence and Interactive Digital Entertainment Conference*	USA, Burlingame, CA	27 May 16	30 Jun 16	10 Oct 16	12 Oct 16	✎
www	**AAAI** *AAAI Conference on Artificial Intelligence (AAAI)*	USA, San Francisco, California	09 Sep 88	11 Nov 16	04 Feb 17	09 Feb 17	✎
www	**hri** *Human-Robot Interaction*	Japan, Tokyo	10 Sep 12		04 Mar 13	06 Mar 13	✎

Figure 2 Screenshot of http://www.confsearch.org/



```
{
"@score":"1",
"@id":"92917",
"info":{"title":"Symbiotic Cognitive Computing through Iteratively Supervised Lexicon Induction.","authors":{"author":["Alfredo Alba","Clemens Drews","Daniel Gruhl","Neal Lewis","Pablo N. Mendes","Meenakshi Nagarajan","Steve Welch","Anni Coden","Ashequl Qadir"]},"venue":"AAAI Workshop - Symbiotic Cognitive
Systems","year":"2016","type":"Conference and Workshop Papers","url":"http://dblp.org/rec/conf/aaai/AlbaDGLMNWCQ16"},
"url":"URL#92917"
},
{
"@score":"1",
"@id":"92918",
"info":{"title":"Measuring Synergy from Benevolence in a Network Organization.","authors":{"author":["Saad Alqithami","Henry Hexmoor"]},"venue":"AAAI Workshop - Multiagent Interaction without Prior Coordination","year":"2016","type":"Conference and Workshop
Papers","url":"http://dblp.org/rec/conf/aaai/AlqithamiH16"},
"url":"URL#92918"
},
{
"@score":"1",
"@id":"92919",
"info":{"title":"Deep Activity Recognition Models with Triaxial Accelerometers.","authors":{"author":["Mohammad Abu Alsheikh","Ahmed Selim","Dusit Niyato","Linda Doyle","Shaowei Lin","Hwee-Pink Tan"]},"venue":"AAAI Workshop - Artificial Intelligence Applied to Assistive Technologies and Smart
Environments","year":"2016","type":"Conference and Workshop Papers","url":"http://dblp.org/rec/conf/aaai/AlsheikhSNDLT16"},
"url":"URL#92919"
},
```

Figure 3 Screenshot of Example Query Result of AAAI (a top conference in Artificial Intelligence field) from DBLP API

Table 9 Computer Science Topics and their Prior Probability

Name	Probability
Artificial Intelligence	0.07883914721691136
Computer Vision	0.16541611052992317
Distributed Computing	0.06739926409581184
Data Mining	0.08138234551423108
Human Computer Interaction	0.0448712167670719
Information Retrieval	0.08127412431008982
Machine Learning	0.12017513798203527
Natural Language Processing	0.04621496338515926
Networks	0.07809963565527939
Operating System	0.0933317701381624
Security	0.007611558024602287
Software	0.030419176797373833
Theory	0.10496554958334836

3.2.2 Results and Issues

Table 10 Computer Science Topics and their representative phrases

Name	SubTopic 1	SubTopic 2	SubTopic 3	SubTopic 4

Artificial Intelligence	solve problem;
	constraint satisfaction;
	logic program
		reinforcement learn;
		large scale;
		neural network
			neural network;
			hidden markov model;
			graphical model
				natural language;
				multi agent system;
				artificial intelligence;

Computer Vision	face recognition;
	support vector machine;
	large scale
		image segmentation;
		optical flow;
		image retrieval
			real time;
			pose estimation;
			super resolution
				video code;
				computer vision;
				neural network

Distributed Computing	query process;
	data stream;
	nearest neighbor
		real time;
		fault tolerant;
		ad hoc network
			web application;
			object orient;
			web service
				high performance;
				large scale;
				load balance

Data Mining	statistical machine translation;
	semi supervise;
	natural language
		data mining;
		time series;
		large scale
			query process;
			database system;
			object orient
				social network;
				question answer;
				social media

Human Computer Interaction	collaborative design;
	web service;
	design system
		activity recognition;
		augment reality;
		real time
			wireless network;
			cognitive radio;
			sensor network
				social network;
				context aware;
				social media

Information Retrieval	question answer;
	query process;
	association rule
		database system;
		data management;
		information system
			large scale;
			social network;
			time series
				information retrieval;
				web search;
				search engine

Machine Learning	support vector machine;
	reinforcement learn;
	semi supervise learn
		data mining;
		feature selection;
		topic model
			problem solve;
			logic program;
			multi agent
				neural network;
				time series;
				real time

Natural Language Processing	question answer;
	relation extraction;
	spoken dialogue
		logic program;
		natural language;
		finite state
			statistical machine translation;
			language model;
			speech recognition
				sentiment analysis;
				word sense;
				name entity

Networks	load balance;
	data center;
	performance analysis
		wireless sensor network;
		energy efficient;
		cognitive radio network
			large scale;
			real time;
			solve problem
				decision support;
				model system

Operating System	high performance;
	load balance;
	distribute system
		low power;
		delta sigma modulator;
		dc converter
			wireless network;
			wireless sensor network;
			fault tolerant
				real time;
				live demonstration;
				video code

Software	logic program;
	higher order;
	answer set program
		garbage collection;
		type inference;
		data structure
			object orient;
			orient program;
			domain specific
				type language;
				parallel program;
				program language

Theory	approximation algorithm;
	shortest path;
	stochastic problem
		context free;
		zero knowledge;
		communication complexity
			stock market;
			stock price;
			stock prediction
				stochastic model;
				stochastic system;
				stochastic network

Table 11 Examples of Query Researchers with Fixed Field

Name	Field	SubTopic	SubSubTopic
Jiawei Han	DM	efficient learning, frequent pattern, time series data, pattern data	frequent pattern,data mining, mining association rule, mining frequent pattern
Fei-Fei Li	CV	convolutional neural network,image classification, event detection, large scale image	visual recognition, spatial temporal, event detection, semi supervise, probabilistic model, visual model, supervise segmentation

Table 12 Examples of Query Researchers without Fixed Field

Name	Classified Field	SubTopic	SubSubTopic
Jiawei Han	IR	matrix factorization, high dimensional, data mining, mining association	mining frequent, data mining
Fei-Fei Li	CV	convolutional neural network,image classification, event detection, large scale image	visual recognition, spatial temporal, event detection, semi supervise, probabilistic model, visual model, supervise segmentation

From the result, we can see that it is very hard to give a clear and understandable meaning for each subtopic for most of the fields. Although we can find some non-overlapping subtopic in computer vision field, for most of the topics the subtopics are indeed highly overlapped with each other, if we try to interpret the main idea of the field with the meaningful phrases. In other words, the depth of the hierarchical tree is hard to define.

For query part, there also exist some problems. For table 11 we can see that if we fix the category and try to query the researcher, the method would generate meaningful results. However, if we use the naive bayes classifier to first classify the category and then query the hierarchical topical tree, the result is not very satisfactory, especially in the field of Information Retrieval, Machine Learning, Natural Language Processing, and Data Mining. The reason is that these fields have so many overlapping conferences, such as Knowledge Discovery and Data Mining (KDD), Conference on Neural Information Processing Systems (NIPS), and International Conference on Machine Learning (ICML). Many titles in these conferences can categorized into many categories, thus the hierarchical topical tree and naive bayes classifier have many overlaps, leading to unsatisfactory results.

4. Conclusion And Future Work

In this thesis, we present two methods for two scenario to generate meaningful and representative phrases for one researcher. The first scenario is to consider each researcher independently, and the proposed method has a satisfactory result. The second scenario is to consider each researcher with respect to another researcher. Our proposed method can work in some case, but would fail in general because the topic model we use is not strong enough, and the dataset is somewhat too hybrid.

For finding the representative phrases for one researcher, we might want to take the user feedback into account when we incorporate the existing system into the website. The user could judge which phrase is not accurate, and also supply the phrase he/she think should include in the five phrases. In this way the result can be improved in a continuous way. For finding the detailed representative phrases or two or more researchers, we might need to find another completely different approaches.

References

[1]: Jialu Liu*, Jingbo Shang*, Chi Wang, Xiang Ren and Jiawei Han, "Mining Quality Phrases from Massive Text Corpora", Proceedings of 2015 SIAM International Conference on Data Mining, Apr. 2015

[2]: Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, Jiawei Han, “A phrase mining framework for recursive construction of a topical hierarchy”, Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Aug. 2013

[3]: Michael Kuhn, ConfSearch, 2015, Available: ‶http://www.confsearch.org/″