Tracking the Relevance of Academic Conferences

Huizi Mao, huizi@stanford.edu

I. Introduction

An academic conference typically span a wide range of topics which changes with time. It is known that topics of conferences in different areas may overlap. Constantly, researchers are interested in the trend of a conference like, what the most popular topic/method/problem in this year's papers is. The relevance of two different conferences reflects such a change as it is an indicator of interdisciplinary research. For example, The wide application of Neural Networks in Computer Vision may lead to increasing relevance of CVPR and NIPS. In this project, I want to propose a measure of relevance and give a quantitative demonstration of how the relevance changes with time.

A paper is simply a document while a conference consisting of a number of papers can be regarded as a corpus. It is interesting how we can find features or representations of documents and further calculate the relevance between them. TF-IDF is a kind of word frequency feature that has been widely used for document representation and similarity matching[1], [2]. Recently with the success of word embedding, there are attempts to map a document direct into vector space(doc2vec[3]).

All papers of a conference form a rich corpus which is suitable for Topic Modeling. Among all possible solutions, Latent Dirichlet allocation(LDA) has been proven an efficient method to estimate the latent topics of a corpus[4]. A number of variants were further proposed, to enable modeling with time[5], authors[6] or even multiple corpora[7]. LDA-based methods give possible relevance metrics based on latent variables or probability distribution. Recently, Word Embedding methods like word2vec[8] also provide possible distance metrics for different topics.

The goal of this project includes the following two aspects: (I) to model or to represent different conferences over different years; (II) to measure the relevance of conferences. In this project we propose four approaches to this problem, and evaluate them both quantitatively and qualitatively.

II. Related Works

Relevance, distance and similarity are nearly equivalent in our study, therefore in the following sections I only use the phrase "relevance" to avoid ambiguity.

A. Distance-based Relevance Measure

The key points of discriminative methods are to find a feature representation and apply supervised or unsupervised metric to it. TF-IDF is an early and well-know method to represent a document[1]. It is applied on word frequency features which are extracted with Bag-of-words(BoW)[9]. Also, the topic distribution extracted by a topic model can also be regarded as a feature and further used for relevance measurement.

Recently with the surge of word embedding related studies, there are attempts to document representations and relevance measurements. Le et al. proposed *doc2vec*, which represents a document with a fixed-length feature vector and outperformed the old BoW feature in many NLP task[3].

Given the features extracted by any kind of methods, there are also a wide range of approaches to measure the distance(or similarity, relevance). The simplest ones are cosine distance, euclidean distance and hamming distance. Kusner et al. integrated word2vec and Mover's Distance and proposed a methods to directly measure the document relevance[10].

B. Probability-based Relevance Measure

The key points of generative relevance measure is to model a given conference in a way that it is able to assign a paper a confidence(probability, in most cases). Naive Bayesian is a simple model the word distribution in a paper with independence assumption.

Topic Modeling, which use "topics" as a latent variable for document representation, is also a generative probability model. Latent Dirichlet Allocation, which is regarded as the most successful topic model, is able to extract the hidden topics in a corpus and the topic distribution over documents in the corpus[4]. Dynamic topic models[5] and Markov topic models[7] further expand it to the situation of time-series topic modeling and multi-corpora topic modeling.

Based on the topics extracted, there are a number of papers trying to construct a distance metrics. Aletras et al. did a survey of relevance measuring methods, compared them with human judgements and concluded that Pointwise Mutual Information(PMI) and Explicit Semantic Analysis(ESA) are potentially the two among the best[11].

C. Previous work on conference relevance study

Wang et al. studied the relevance of conferences back in 2009 in their work of Markov Topic Model[7]. They studied the relevance of six conferences CIKM, ICML, KDD, NIPS, SIGIR and WWW from 2005 to 2008. However, they did not pay attention to the time-series information, i.e., the change of relevance over time, which is the focus of this project.

III. Dataset

The dataset were crawled from their websites using a crawler written by myself. It contains five conferences (AAAI, AISTATS, AIIDE, NIPS, ACL) from 2010 to 2016. For each paper, the crawler only stored four fields: title, authors, abstract and full text. A brief introduction of these five conferences are given here.

- **AAAI** AAAI Conference on Artificial Intelligence
- **AISTATS** Artificial Intelligence and Statistics Conference
- **AIIDE** Artificial Intelligence and Interactive Digital Entertainment
- **NIPS** Neural Information Processing Systems
- **ACL** The annual meeting of the Association for Computational Linguistics

In the initial plan, we only evaluate the method on abstracts. However, problems emerge with only the abstracts. First, conferences like AAAI usually have only 1-2 sentences for the abstract. Second, some conferences like ACL do not highlight their abstracts in the website, which makes harder crawling only the abstracts. Due to these reasons, we evaluate out methods on full texts, which are extracted from the pdf file.

The basic statistics of these five conferences are listed as Table III. The visualization is given in Figure 1. There are 9301 papers in total.

	2010	2011	2012	2013	2014	2015	2016
AAAI	346	342	382	277	474	673	690
AISTATS	125	99	157	71	122	126	164
AIIDE	43	36	33	40	35	34	37
NIPS	292	306	368	360	411	402	563
ACL	269	344	222	394	328	358	378

TABLE I
EVER-YEAR NUMBER OF ACCEPTED PAPERS

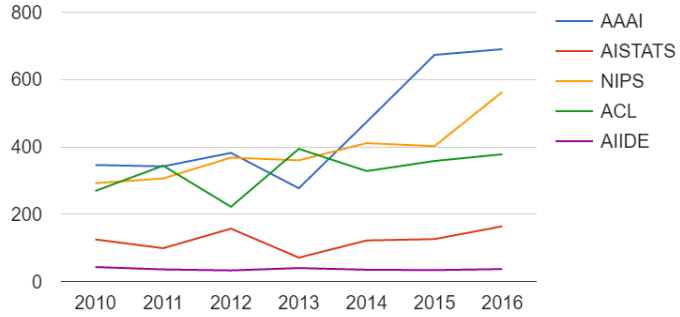

Fig. 1. Number of papers in every conference over year

IV. METHODS

I propose four possible solutions to conference prediction. The first two, TFIDF-Cos and TM-Cos are basically distance-based while the latter two are probability-based.

TF-IDF and cosine similarity(TFIDF-Cos) It calculates the word frequency of every paper and apply TF-IDF to extract the feature vectors of every document. The feature vector of a conference is the average of all its papers. The similarity is then calculated with simple cosine distance.

A mathematical explanation of TF-IDF is given as Equation 1. The formulation of TF-IDF is referred to Wikipedia[1]. Our algorithm is shown in Equation 2. t is the token.

[1]Wikipedia:TF-IDF

$$\begin{aligned} \mathrm{tf}(t,d) &= 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}} \\ \mathrm{idf}(t,D) &= \log \frac{N}{|\{d \in D : t \in d\}|} \\ \mathrm{tfidf}(t,d) &= \mathrm{tf}(t,d)/\mathrm{idf}(t,D) \end{aligned} \tag{1}$$

$$\begin{aligned} \mathbf{f}_t(d) &= \mathrm{tfidf}(t,d) \\ \mathbf{f}(C) &= \frac{1}{|C|}\sum_{d\in C}\mathbf{f}(d) \\ \mathrm{Similarity}_{C_1,C_2} &= \mathbf{f}(C_1)\cdot\mathbf{f}(C_2)/|\mathbf{f}(C_1)||\mathbf{f}(C_2)| \end{aligned} \tag{2}$$

Topic-model-based Cosine Similarity(TM-Cos) It treats all conferences in a year as one big corpus and apply topic modeling to it. Topic model is able to model the topic distribution(or composition) of a paper. Intuitively, papers with similar topic distributions are more relevant, thus we can measure the relevance based on Cosine distance of their topic probability. The topics of a conference is the average of all papers' latent topic distribution. The similarity of a conference and a paper is based on the cosine similarity of the latent topic variables, with no consideration about the time-series information.

To be specific, the classifying process is identical to the baseline method except that the feature is the probability of topic distribution of a document as indicated in Equation 3. The feature is chosen as θ_d, which is the topic distribution of document d.

$$\begin{aligned} \mathbf{f}(d) &= \theta_d \\ \mathbf{f}(C) &= \frac{1}{|C|}\sum_{d\in C}\mathbf{f}(d) \\ \mathrm{Similarity}_{C_1,C_2} &= \mathbf{f}(C_1)\cdot\mathbf{f}(C_2)/|\mathbf{f}(C_1)||\mathbf{f}(C_2)| \end{aligned} \tag{3}$$

Naive Bayesian similarity(NB)Naive Bayes model is a simple probabilistic model based on applying Bayes' theorem with independence assumptions between the words. Consider the existence of some unique words in some papers, Laplacian smoothing is also adopted.

As shown in Equation 4, given a paper with a word histogram of $H_w(d)$ and the word frequency $F_{c,w}$ of conference c, we can obtain the log-probability of paper d in that conference. Here $F_{c,w}$ has already been smoothed. Notice that we ignore the bias term here, which is the prior conference probability, because it is based on the number of papers in a conference.

$$\begin{aligned} \mathbf{L}_c(w) &= \sum_{winW} log(F_{c,w})H_w(d) \\ \mathrm{Similarity}_{C_1,C_2} &= \sum_{winC_1}\mathbf{L}_{C_2}(w) \end{aligned} \tag{4}$$

Generative Topic Model similarity(GTM) LDA is able to model the topic distribution of a document and further calculate the probability of that document. As shown in Equation 5, given a paper with a word histogram of $H_w(d)$, topic distribution $\theta_c(d)$ and word distribution of all topics $\psi_c(t)$, we are interested in the log-probability of document

d. Notice that we use an approximate representation for the consideration of simplicity.

$$\begin{aligned}\mathbf{L}_c(w) &= log(\sum_{t=1}^{T}\theta_{c,t}(d)\prod_{winW}\psi_{c,w}(t)^{H_w(d)}) \\ &\approx max_{t=1}^{T} log(\theta_{c,t}(d)) \\ &\quad + \sum_{winW} log(\psi_{c,w}(t))H_w(d) \\ \text{Similarity}_{C_1,C_2} &= \sum_{winC_1}\mathbf{L}_{C_2}(w)\end{aligned} \tag{5}$$

A. Data preprocessing

Data preprocessing is necessary in most all NLP tasks. Common techniques are adopted in this project, including Tokenization, Elimination of stop words and Stemming. I follow the instruction of the online tutorial[2]. Besides common stop words, I also add additional words which are common in this area but actually meaningless. These words are 'use','problem','can','method','approach', 'show','result', 'model'.

Another problem is noise, two types of noise one coming from the extraction process from pdf files, the other caused by misspelling words, abbreviation and human names. For the first type of noise, a simple solution is to delete tokens that are too short or contain special characters. For the second case, however, it is much more complicated. One solution (or circumvention) I came up with is to delete the token that only occur once in the article. It is based on the intuition that special words usually occur only once while meaningful words occur more frequently. Empirical study of this strategy is given in the experiment section.

V. Experiements

A. Evaluation Metric

The methods will be evaluated from two perspectives: how well the method models the corpora, and how well the relevance is measured.

It is easy to evaluate the effectiveness of the model by intuition. For example, a researcher may notice a growing connection between Computer Vision, Natural Language Processing and Deep Learning by his experience. In the work Markov Topic Model[7], Wang et al. gives a qualitative analysis based on visualization and human visualization. Thus, I will present the topic extracted from every conference and compare it to the visualized relevance curve, in order to find out whether the results make sense.

However, it is quite difficult to quantitatively measure the correctness of relevance. To better address this problem, some criterion is required to measure how the relevance are modeled. Without doubt a conference is most similar with itself. Based on that fact, we intuitively make an assumption that, the similarity between one conference in two following years should be higher than two different conferences. Based on this assumption we propose a metric, Relative Variance(RV), to evaluate the accuracy of the model. Denote by $C_{i,t}$ the i-th conference in year t where $i = 1, 2, \cdots, n$, $t = 1, 2, \cdots, T$. $f(c_1, c_2)$ is any function measure the similarity between c_1 and c_2 and $f(c_1, c_2) \in [0, 1]$. RV is formulated as:

$$\begin{aligned}RV &= W_{t,n}\frac{\sum_{t=1}^{T}\sum_{i\neq j}^{n} f(C_{i,t}, C_{j,t})}{\sum_{t=1}^{T-1}\sum_{i=1}^{n} f(C_{i,t}, C_{i,t+1})} \\ W_{T,n} &= \frac{n}{n(n-1)}\frac{T-1}{T}\end{aligned} \tag{6}$$

In Equation 6, $W_{t,n}$ is some constant only dependent on T and n. Our aim is to minimize RV.

B. Experimental Results

The Topic model and TFIDF implementation are dependent on gensim[12]. For TM-Cos and GTM, the number of topics is chosen as 10. All topic models are trained for 20 passes with a batch size of 2000.

Quantitative results of RV metric is given in Table V-B. From the statistics we find that distance-based methods generally give a better results. TM-Cos has the lowest RV score of 0.392.

TABLE II
A DETAILED COMPARISON OF ALL METHODS

	TFIDF-Cos	TM-Cos	GTM	NB
RV	0.498	0.392	0.951	0.858

Visualizations of the correlation between AAAI and other conferences is given in Fig. 2. Though TM-Cos method has a lower RV score, TF-IDF gives smoother curves and more making-sense results. It may be due to that topic features are not suitable for direct distance computing. Another important factor is that, though it indicates whether the relevance is reliable, RV does not put any constraint over the time consistency of the relevance metric. The problem of smoothness also occurs to GTM and NB methods.

From the figure given by TFIDF-Cos, we can find NIPS always had a strong correlation with AAAI. ACL and AIIDE both had a weak correlation between AAAI, but as time went ACL was getting a constantly stronger correlation.

C. Tuning the methods

We also want to know how the "word filtering trick", which is described in Section IV-A, influence the results. Table V-C shows the difference between with/without word filtering preprocessing. Though it lowers the RV metric for some methods, the overall improvements is negligible. Therefore we do not apply it to our final results.

TABLE III
COMPARISON OF MEAN PRECISION WITH/WITHOUT WORD FILTERING PREPROCESSING

RV	TFIDF-Cos	TM-Cos	GTM	NB
without dp	0.498	0.392	0.951	0.858
with dp	0.484	0.411	0.946	0.894

[2]Latent Dirichlet Allocation (LDA) with Python

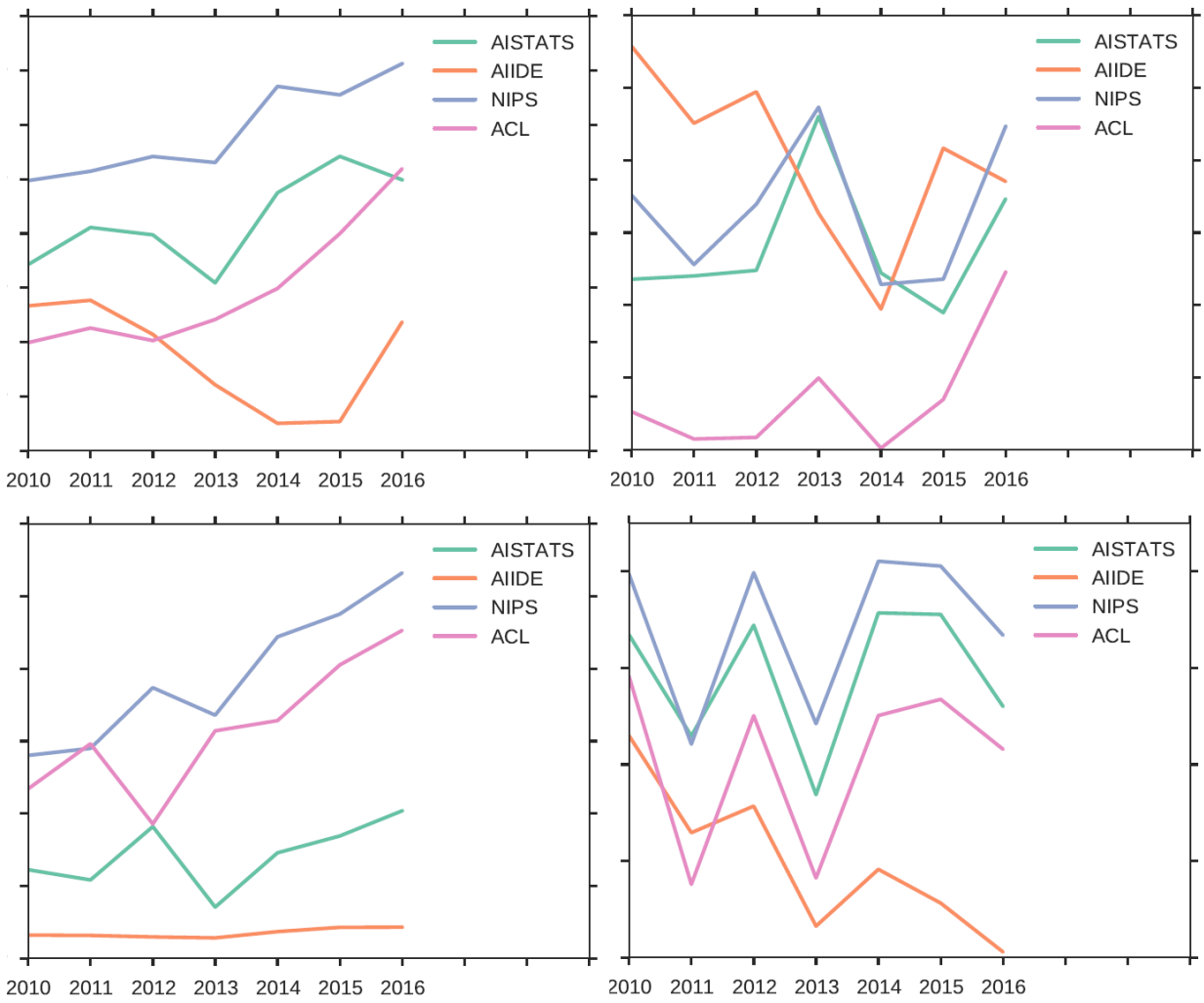

Fig. 2. Time-series correlation between AAAI and other conferences. Top-left:TFIDF-Cos; Top-right:TM-Cos; Bottom-left:GTM; Bottom-right:NB

D. *Insights from topic model*

To give an intuitive understanding of how these conferences are correlated and how difficult to distinguish them, we first use LDA topic model to model the topic distribution of every article in year 2014. The topic number is set to 10. After that, we randomly choose 35 papers from every conference and construct a 165×10 matrix. To visualize those vectors we apply PCA to compress it to 165×2.

The result is shown in Fig.3. As we can see from the figure, points of ACL and AIIDE are far away from those pf the other three conferences, therefore easy to classify. Such a distribution matches the correlation in Figure 2.

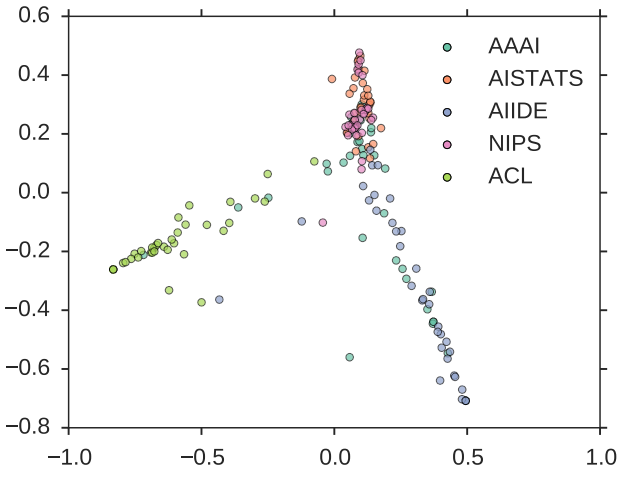

Fig. 3. Visualization of topic distribution after PCA

VI. Conclusion and Future Works

In this project, we constructed a dataset of 5 conferences with full texts of 9301 papers from 2010 to 2016. To measure the relevance between those conferences over years, four methods were proposed. Those methods are evaluated empirically by visualization. In addition, we proposed Relative Variance as a metric to evaluate how well those methods model the relevance. A remaining problem is how to get a metric to evaluate time-series correlation, and further to propose a method that measures the relevance with a time-consistent

manner.

REFERENCES

[1] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[2] Juan Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 2003.

[3] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[6] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[7] Chong Wang, Bo Thiesson, Christopher Meek, and David M Blei. Markov topic models. In *AISTATS*, pages 583–590, 2009.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[9] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[10] Matt J Kusner, Yu Sun, Nicholas I Kolkin, and Kilian Q Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 957–966, 2015.

[11] Nikolaos Aletras and Mark Stevenson. Measuring the similarity between automatically generated topics. In *EACL*, pages 22–27, 2014.

[12] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.