Published Online 15-March-2021

FINDING OUT THE INFLUENCE OF TOPIC BASED OBJECT ACCORDING TO SOCIAL NETWORKS

Huynh Tan Hoi
Faculty of Business Administration, Ho Chi Minh City Open University
Corresponding Author: hoi.ht0220@oude.edu.vn

Abstract: ***In this research paper, the author mainly focuses on researching models in social networks analysis and then use them to develop the system which is able to allow users to search right topics as well as right research papers of other authors. In this paper, the author do research on Author – Conference –Topic model for the topic relevant to users' requirements. With each of these topics, the system will create a graph of Topical Affinity Propagation (TAP) with the aim to analyze the topics based influence of authors. The results on the corpus extracted from Microsoft Academic Research include 34,330 authors, 19,921 articles, 1,335 conference and 424,501 co-author relations.***

Keyword: requirements, analysis, social networks, topics.

Introduction

The speed of Internet development has enabled people to access information in the world quickly. However, with a huge amount of information and constantly growing, there have appeared many effective search support tools such as Google, Bing, and Yahoo! Search. These tools somewhat meet the needs of users, but the results returned sometimes do not match the desired. For example, when typing a keyword "Data Mining" not only desires to find documents containing this word if we also want to find related topics in the same field. The next problem, after finding relevant topics is how to identify experts for each topic. Aiming to create tools to support learning, research the scientific works of leading experts, this study uses the ACT model [2] to explore TAP topics and models, to analyze the influence of each author to identify experts on each topic. The rest of the paper is organized as follows: Part 2 (Author Model - Conference – Topic); Part 3 (Graph of spreading the relationship by topic); Part 4 (Test results and discussion); Part 5 (Conclusion and development direction).

Author-Conference-Topic Model

Originating from the Latent Dirichlet Allocation model proposed by Blei, Ng and Jordan in 2003 following the Bayes network approach, the LDA model is used to model the corpus to discover hidden topics of the corpus [3]. The process of generating a document consists of three steps: (i) for each document that has a subject probability distribution of that document, this distribution is sampled from the Dirichlet probability distribution, (ii) for each word in the document, a unique topic is chosen from the above topic distribution, (iii) each keyword will be drawn from the polynomial distribution for the keyword according to the chosen topic. Many models have been developed from LDA model such as Author-Topic [4], AuthorRecipient-Topic [1], particularly the Author-Conference-Topic model [5] is an improvement model from the Author-Topic model, in addition to a topic arising from a specific word set, it also allows a topic to generate relevant conferences. Some related concepts are as below:

A_d : authors writes documents d

θ_x : probability distribution topic of X authors who are interested, this distribution satisfies Dirichlet distribution.

φ_z : probability distribution of keywords generated by a topic z, this distribution satisfies the Dirichlet distribution.

W_{di} : a set of keywords selected from the Multinomial polynomial distribution (φ_z)

α: parameter focusing for subject probability distribution

β: centralized parameter for the characteristic magnetic probability distribution

c_{di} : conference set selected from Multinomial polynomial distribution (φ_z)

μ: centralized parameter for conference probability distribution

ψ_z : conference file (conference, article) belongs to topic z

w_{di} : the ith keyword that is generated by the topic

zdi is selected from polynomial distribution (φ_z)

z_{di} : subject z assigned to the i-th keyword generated by the author of xdi chose the polynomial distribution (θx)

x_{di} : the author is chosen for each i th word in document d generated by the uniform distribution (ad)

The generation process of the ACT model is as follows:
For each z topic, the probability distribution for the keywords φ_z and conferences ψ_z will be calculated according to the Dirichlet probability distribution (β) and the Dirichlet probability distribution (μ).
For each wdi keyword in the article d:

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Select an xdi author from the author distribution set ad, this distribution satisfies the uniform distribution (a_d).

Select a zdi topic from the subject distribution of the xdi author, which satisfies the polynomial distribution (θ_{xdi}).

Where θ is the topic probability distribution Dirichlet (α).

Choose a wdi keyword from the distribution of the zdi topic, this distribution satisfies the polynomial distribution (ψ_{zdi}).

Select a cdi conference from the conference distribution of zdi topic, this distribution satisfies the polynomial distribution (ψ_{zdi}).

Similar to the LDA model, the ACT model needs to estimate two parameters: (1) the distribution θ is the author-topic matrix, the distribution φ is the subject-keyword matrix (Topic). -Words), distribution ψ is the topic-conference matrix (TopicConferences); The parameter (2) is the probability that the topic zdi and the author xdi are assigned to the word wdi. We will then use Gibbs sampling [11] to estimate the parameters for the two variables x and z as follows:

$$P(z_{di}, x_{di} \mid z_{-di}, x_{-di}, w, c, \alpha, \beta, \mu) \propto \frac{m^{-di}_{x_{di}z_{di}} + \alpha}{\sum_z (m^{-di}_{x_{di}z} + \alpha)} \frac{n^{-di}_{z_{di}w_{di}} + \beta}{\sum_v (n^{-di}_{z_{di}v} + \beta)} \frac{n^{-di}_{z_{di}c_{di}} + \mu}{\sum_c (n^{-di}_{z_{di}c} + \mu)} \qquad (1)$$

During implementation, 3 matrices A x T (author - topic), T x V (topic - keywords) and T x C (topic - conference) will be calculated. From these 3 matrices, the probability of generating a topic z by an author x is φ_{zv}, the probability of generating a word v by a topic z is φzv, and the probability that a conference c is generated by a leader thread z is ψ_{zc} according to the following Gibbs [11]:

$$\theta_{xz} = \frac{m_{xz} + \alpha}{\sum_{z'} (m_{xz'} + \alpha)} \qquad (1)$$

$$\phi_{zv} = \frac{n_{zv} + \beta}{\sum_{v'} (n_{zv'} + \beta)} \qquad (2)$$

$$\psi_{zc} = \frac{n_{zc} + \mu}{\sum_{c'} (n_{zc'} + \mu)} \qquad (3)$$

Where m_{xz} is the number of times author x has arisen topic z, mxz 'is the number of times author x has arisen topic other than z (z ≠ z ’). And nzv is the number of times z topic arises from characteristic v, nzv 'is the number of times z topic arises from characteristic v differs from v (v ≠ v ’). Lastly, nzc is the number of times z has raised a conference c and $n_{zc'}$ is the number of times the host has received the topic so that z arises another conference c (c ≠ c ’).

The Model

TopicalAffinity Propagation model proposed by Jie Tang, Jimeng, Chi Wang and Zi Yang [6] in 2009, using factor graph model [4] in combination with affinity propagation model [2] to analyze the influence of each node in the graph. However, different from the two algorithms used to analyze each node's influence in the graph is the separate Vector algorithm and PageRank, the TAP (Topical Affinity Propagation) model [6] can give a specific level of possible influence between buttons. In order to analyze the influence in the co-authoring network, the following important information for research is needed.

A set of relationships between authors in the network

Number of articles the author writes for topics they are interested in.

Probabilistic distribution by subject of an author x.

For example, the author "Jiawen Han" is interested in 3 topics with a probability distribution of {"Data Mining" => 0.642857, "Database" => 0.285714, "Information Retrieval" => 0.071429}

- Topical influence: the influence of s button to t button (symbol μ_{st}) including two variables r (t, s) and a (t, s).

- Building a graph G = (E, V) where V is a set of nodes (the authors), E is a set of edges representing the link between authors, each has a weight corresponding to the number of articles that the two authors participated in. A set of topic probability distributions for each participating node in G. From these data, formalize the impact analysis into the Topical Factor Graph [6] as follows: Give a set of nodes to be analyzed $\{v_i\}_{i=1}^{N}$,, where N is the number of nodes on the network and a hidden vector set $\{y_i\}_{i=1}^{N}$is used to represent nodes with The probability that interest in the highest topic affects the node v_i and z y_i is the button that has the most influence on topic z on node v_i.

Local property function of each node [6]: the effect of each author on each topic is calculated by the following formula:

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

$$g(v_i, y_i, z) = \begin{cases} \dfrac{w^z_{iy_i^z}}{\sum_{j\in NB(i)} (w^z_{ij} + w^z_{ji})} & y_i^z \neq i \\ \dfrac{\sum_{j\in NB(i)} w^z_{ji}}{\sum_{j\in NB(i)} (w^z_{ij} + w^z_{ji})} & y_i^z = i \end{cases} \qquad (4)$$

Where vi is the node considered in topic z, y_i is the node with the highest probability of affecting the topic z on the node vi. . $w_{ij}^z = y_j^z \cdot \propto_{ij}$ with y_j^z is the highest probability of influence in topic z on node vi and αij is the weight equivalent to the number of articles between v_i. and *vj* nodes (shared by 2 authors)). And j ∈ NB (i) are adjacent nodes vi. In case the node vi under review is also the most influential node in topic *z* $(y_i^z = i)$,we calculate the influence of v_i. to the nodes adjacent to it vj ∈ NB (i), otherwise $(y_i^z \neq i)$ and simply calculate the effect of the button r_{ij}^z directly on v_i. itself.

The set of each node in formula (5) only calculates the influence of the object on the topic. Next we will calculate the influence score among the nodes in the graph. In the co-authoring network, r_{ij}^z is information exchanged from node i to node j according to topic *z*, a_{ij}^z is information sent from node *j* to node i initialized to zero. measure the influence of each pair of authors on each topic:

$$b^z_{ij} = \log \frac{g(v_i, y_i, z) \mid y_i^z = j}{\sum_{k\in NB(i)\bigcup\{i\}} g(v_i, y_i, z) \mid y_i^z = k} \qquad (6)$$

Formula (6) normalizes the impact of each pair of authors on each topic, the log function is used to switch from the sum-product calculation (SumProduct) to the Max-Sum calculated by the SumProduct algorithm when applied on graphs factor, the value of the functions will be too small (decimal numbers contain many zeros) [10]. The TAP algorithm [6] extends the propagation algorithm according to the relation [2] whose input consists of three variables, b_{ij}^z is the degree of influence of node i and j button in topic *z*, variable r_{ij}^r is node i is affected by node *j*, variable a_{ij}^z is the degree of influence *j* button on node *i*. Function of exchanging information from node i to node j [6]

$$r^z_{ij} = b^z_{ij} - \max_{k\in NB(j)} \{b^z_{ik} + a^z_{ik}\} \qquad (7)$$

Formula (7) describes node i sending information to node j to determine the influence of node i by node j by denoting the degree of influence of node *i* and node j in subject *z*, b_{ij}^z , with the maximum value of nodes k adjacent to node j, node i is also affected by Frey and Dueck [2], if the value of the influence of node *i* by node *j* is positive, then proof show that j button has influence on *i* button. The variable a_{ik}^z sends information from node *k* (adjacent to node *j*) to node i to reflect the degree of influence of node *k* on node *i*, initially starting at zero.

Function of exchanging information from node j to node i [6]

$$a^z_{ij} = \min(\max\{r^z_{jj}, 0\}, -\min\{r^z_{jj}, 0\} - \max_{k\in NB(j)\backslash\{i\}} \min\{r^z_{kj}, 0\}), i \in NB(j) \qquad (8)$$

Equation (8) describes node j sending information to node i to determine the degree of influence of node *j* on node i. The variable r_{ij}^z indicates the degree of influence *j* itself on the *z* topic. Because the propagation algorithm in relation is clustering algorithm, Frey and Dueck [2] think that if the value of each data point is too large, each point itself is likely to become a separate cluster, whereas if the price If each data point is too small, all of these points tend to gather in the same cluster.

According to Frey and Dueck, the best value of node j affecting node i is within the smallest and largest values of node j, of which the largest value is the degree of influence j node itself for topic z determined by the function $\max\{r_{ij}^r, 0\}$, and the smallest value is the sum of the smallest values of the degree of self-influence of j with the largest value in the nodes k adjacent to j under the influence of node j is determined as $-\min\{r_{jj}^z, 0\} - \max_{k\in NB(j)\backslash i} \min\{r_{kj}^z, 0\}$. Because j is

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

considering the influence of j on the node i, so only select the effect value of the k nodes adjacent to j with negative values by the function $\min_{k \in NB(j) \setminus i} \{r_{kj}^{z}, 0\}$. The min function of the whole formula avoids the fact that the value of a_{ij}^{z} is so large that each node becomes a separate cluster [2]. Influence score function [6]

$$\mu_{st}^{z} = \frac{1}{1+e^{-(r_{ts}^{z}+a_{ts}^{z})}} \quad (9)$$

This sigmoid function converts the values of two variables r_{ts}^{z}, a_{ts}^{z} , a into probability.

Test Results

Data Collection

The first is to have information author related to information technology. Currently, almost all scientists are working at the top universities in the world, each author publishes personal information on the Web. There are many Web sites that have gathered author information as well as scientific papers such as the DBLP, which provides information about the article, or such as CiteseerX, which is a document and author search system. Currently, Microsoft has launched the Academic Search Web site that provides quite enough information about authors, topics, hobbies, conferences, scientific articles, etc. Therefore, we will select this Web site to collect data forming the co-author network.

The data selected and extracted include 34,330 authors, 19,921 articles, 1,335 conferences, 424,501 co-operative relationships (taken from the Microsoft Academic Research website provided by Microsoft at the website address http: // academic.research.microsoft.com/?SearchDomain=2 and then fine-tuned to the program system). From this data, we will build a temporary system called the information technology academic system to search for topics and experts.

Find Topics Related to Keywords

We will use the ACT model [5] as appropriate to the system built, to model all information technology documents. After applying the ACT model, we get the T x V matrix, which indicates the probability that a keyword v ∈ V is generated by the subject t ∈ T. From the above matrix easily identify topics related to the keyword you are looking for. How to do the following: Call the input keyword q (here q is a double word or more), first remove the words that often occur, often called stopword.

Then will break q into single words {q1, q2,. . ., qn} to perform the search. As a result, a topic with multiple keywords appears to match q. From here, we just add the probability of each keyword by each topic and arrange them in descending order. Finally, we choose 5 topics with the highest probability of returning results to users.

Identify Experts by Topic

The identification of experts by each topic is the analysis of the influence of author a on the remaining authors related to a. In this study, TAP model [6] will be used because of its running time and higher accuracy than the separate vector algorithm and PageRank. Other

For traditional graphs that only include nodes, edges and weights, this model is a combination of traditional graphs and statistical probabilities. In order to calculate the impact of author a on author b under the topic z, μ_{ba}^{z}, TAP will consider the influence level with nodes adjacent to node b, then, again, TAP again considers the influence of nodes adjacent to node a to determine the influence of node b on a, μ_{ba}^{z} .In addition, TAP must rely on calculating the influence of the author with the highest probability of interest in the topic being considered. After taking turns calculating the influence between the pairs of nodes in the graph, the next thing is to add the probability by each node to give its influence by topic.

The system includes the following components

Archive component: this is a component for storing co-author social network data, including author profiles, scientific articles, conferences based on the ACT model approach. Subject search component: This component is responsible for receiving requests from users, then the system will find topics related to the keyword that the user has provided. For example, if users enter the keyword "Data Mining", the system will return related topics such as "Database", Information Retrieval ". For this component, use the ACT model to solve. From the resulting topics, users can choose a topic to know which authors are most influential [12].

Components of influence analysis

This component performs analysis based on each topic, by applying a graph model combining statistical probability and TAP algorithm, to assess the influence of the authors as well as find the author. Finally, the composition indicates that the central author set has relationships with the most authors.

Test Results and Evaluation

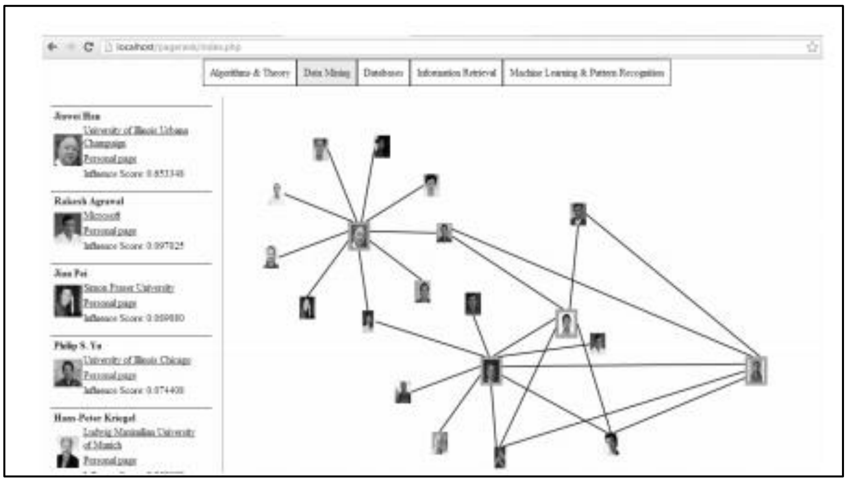
Applying the ACT model for testing on co-authoring network of 34,330 authors, 19,921 articles, 1,335 conferences and 5,258 vocabulary for information technology to generate 24 topics and be labeled manually.

We have installed three separate Vector algorithms, PageRank and TAP on the theme of "Data Mining" including 2,825 authors and 40,110 relationships on the same computer with dual core i3 CPU configuration 2.4GHz, 2GB RAM. The evaluation of the accuracy of these three algorithms will be done as follows: select a list of the top 10 authors in the field of "Data Mining" from the Microsoft Academic

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Research Web site as the standard for evaluating the accuracy of analytical algorithms [13]. For each result from the three algorithms will continue to count the authors that appear in the sample, then divide by the total number of authors. From Table 5 shows that the TAP algorithm has faster runtime and higher accuracy than the other 2 algorithms. Therefore, we choose the TAP algorithm for the program system.

Test Program


localhost/pagerank/index.php
Algorithms & Theory
Data Mining
Databases
Information Retrieval
Machine Learning & Pattern Recognition

Jiawei Han
University of Illinois Urbana Champaign
Personal page
Influence Score: 0.853348

Rakesh Agrawal
Microsoft
Personal page
Influence Score: 0.097825

Jian Pei
Simon Fraser University
Personal page
Influence Score: 0.069880

Philip S. Yu
University of Illinois Chicago
Personal page
Influence Score: 0.074408

Hans-Peter Kriegel
Ludwig Maximilian University of Munich
Personal page

Figure.1. Illustrating the operation of the system when the user types the keyword "Data Mining", the system returns 5 related topics

.

The user will then select a topic to find experts in that topic. In this example, a list of experts on the subject of "Data Mining" is organized by impact score (Figure 1)

Conclusion

In this paper, two main issues are presented to build an academic research support system through the scientific works of leading information technology authors. The first problem is finding topics related to the keywords that users provide by using the Author-Conferences-Topic model, the second problem is finding experts in each topic by using the Topical Affinity Propagation to analyze the impact of each author on the topic they are interested in. The topic finder and object impact analysis system is an illustrative example of social network usage. However, this system still has many improvements as follows: (1) Understanding parallel and distributed algorithms in order to process the growing data of co-author social networks; (2) The analytical model only applies to homogeneous, scalar and weighted networks. Between the two nodes, they do not distinguish their roles (such as an article consisting of a facilitator, a research student or two authors who work together.)

Acknowledgement

The author would like to express warm thanks to the financial support of Ho Chi Minh City Open University, Vietnam.

Ethical clearance

The authors ensure the quality and integrity of the research. By writing this research paper, the author surely read related materials and books.

Conflicts of interest

No conflicts of interest noted in the paper.

References

1. McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2005). The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. Computer Science Department Faculty Publication Series, 44.
2. Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. science, 315(5814), 972-976.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
4. Kschischang, F. R., Frey, B. J., & Loeliger, H. A. (2001). Factor graphs and the sum-product algorithm. IEEE Transactions on information theory, 47(2), 498-519.
5. Tang, J., Jin, R., & Zhang, J. (2008, December). A topic modeling approach and its integration into the random walk framework for academic search. In 2008 Eighth IEEE International Conference on Data Mining (pp. 1055-1060). IEEE.
6. Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2012). The author-topic model

for authors and documents. arXiv preprint arXiv:1207.4169.

7. Jie Tang, Jimeng Sun, Chi Wang, Zi Yang. (2009). Social Network Analysis in large-scale
8. networks, KDD 2009, 807 – 816.
9. Hoi, H. T. (2019). Using Social Networks for English Teaching and Learning. In Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference (pp. 173-177).
10. Darling, W. M. (2011). A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 642-647).
11. Phan, T. H. V., Ho, T. T., & Do, P. (2013). Topic based object Influence analysis in Social networks. Science and Technology Development Journal, 16(4), 68-78.
12. Zhao, Z., Feng, S., Wang, Q., Huang, J. Z., Williams, G. J., & Fan, J. (2012). Topic oriented community detection through social objects and link analysis in social networks. Knowledge-Based Systems, 26, 164-173.
13. Peng, S., Wang, G., & Xie, D. (2016). Social influence analysis in social networking big data: Opportunities and challenges. IEEE network, 31(1), 11-17.
14. Kim, H. L., Breslin, J. G., Chao, H. C., & Shu, L. (2011). Evolution of social networks based on tagging practices. IEEE Transactions on Services Computing, 6(2), 252-261.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731