A Personalized People Recommender System Using Global Search Approach

Chun-Hua Tsai, Peter Brusilovsky
University of Pittsburgh, Pittsburgh

Abstract
The goal of people recommender system is to generate meaningful social suggestion to users. The abundant data are the key factor in fulfilling a recommendation task, but the cost of user data in a real-world system is high. In this paper, we propose a novel approach that integrates a global search result with a personalized people recommendation system. Our approach utilizes the user identity as a query keyword and processes the search results through five different customized parsers. This approach solves the cold-start issue in recommendation systems and leverages the cross-domain information in order to provide a better recommendation result. To test our approach, we embedded it into an existing conference navigator system then deployed the system at two international conferences. The survey results indicate largely positive feedback about the system's effectiveness. Our study results also shed some light on the social interactions that take place at an academic conference.
Keywords: cold-start; personalized; people recommendation system
doi: 10.9776/16601

Contact: cht77@pitt.edu

1 Introduction

Scholars attend academic conferences to expose their work and make a meaningful social connection with other researchers. However, for junior scholars or newcomers to the community, it is hard to establish an effective connection with other attendees. A personalized social support system may help attendees to better find people that fit their research interests or social preferences. It is challenging to fulfill a recommendation task based on limited new user profile data. The cold-start recommendation is always a potential research subject for any system lacking abundant data. To solve this issue, previous studies suggested we can 1) collects more data from users, or 2) adopt collaborative or content-based filtering for cross domain user modeling (?, ?, ?).

There is a straightforward way to explore a person of interest: Google them. Often, we can easily find the publicly search-able information about him or her via Google, but in some cases, Google isn't particularly insightful. Since most conference attendees are academics, and the key to success in academia is to publish your work (?, ?), this implies academics are likely more willing to expose their resume, publication, and social media accounts in a public manner. Hence, we can leverage this understanding to actively collect various user data and profiles from the search engine. This idea corresponds to the two main approaches that we learned from previous studies on the matter.

In this work-in-progress paper, we first examine the effectiveness of the people recommendation system in an academic conference [1]. We propose a novel approach to collect user data from a global search basis. This approach helps us to actively retrieve heterogeneous data for new users. Moreover, we can also adopt existing cross-domain recommendation methods to utilize the data from multiple sources (?, ?). We build up a system that queries user information from Google Search Engine by user identity keywords. Based on the search result, the system can retrieve the data from multiple web sources and generate the people recommendation through the rich information. We then proceed survey in two conferences that the majority user feedback indicates the effectiveness of this system.

[1] The actual system please browser: http://halley.exp.sis.pitt.edu/cn3/portalindex.php

2 Approach

2.1 System Structure

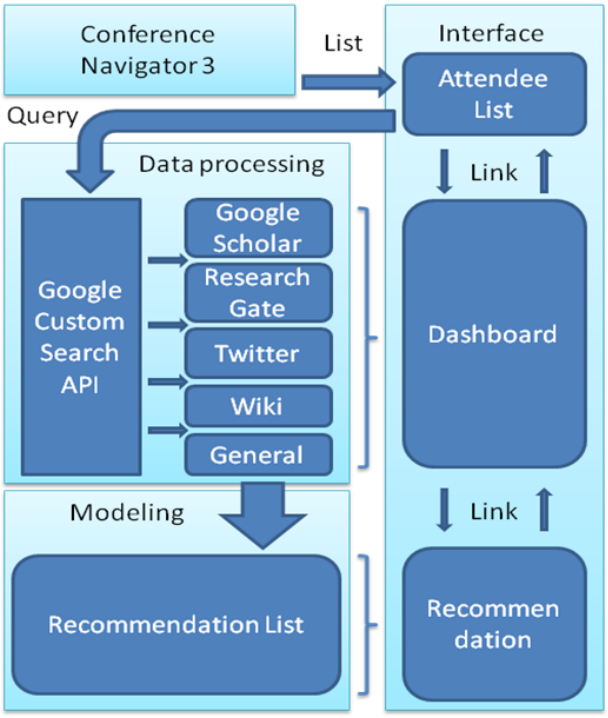

Figure 1: Overview of the proposed system structure.

The goal of this system is to make a people recommendation for the conference attendees. We fetched the conference attendees' list from the Conference Navigator 3 system (?, ?). The list included the attendees' names, affiliations and their current paper titles. We then send the name and affiliation as the query term to Google Custom Search API [2]. Based on the returned search results, we could retrieve further information from additional data sources (e.g. Google Scholar[3], Researchgate[4], Twitter[5], Wikipedia[6] and the other general search result) if needed. After retrieving the data, we need to determine a personalized recommendation score based on these data. The score that is used to make the recommendation is linear, and is combined with content and network similarity. The overview of the system structure is shown in Figure 1.

2.2 Data Processing

Each query fetches up to 100 search results. The downloaded web pages will be parsed by the data sources accordingly. For example, we can parse the name, affiliation, co-author list, publication list and the paper abstracts from the Google Scholar. We can further fetch the tweets and following/follower network from Twitter, and the knowledge-based introduction from Wikipedia. From the HTML pages, we generally extract the main text from each page. It doesn't guarantee to fetch all the web services for each query (i.e. The author might not have a Wikipedia page). Furthermore, if the search result is duplicated, the system will pick the highest ranking result.

[2]https://www.googleapis.com/customsearch/v1
[3]https://scholar.google.com/
[4]http://www.researchgate.net/
[5]https://twitter.com/
[6]https://en.wikipedia.org/

2.3 Model

The system will use the parsed data to calculate a recommendation score for each attendee at a conference. The total score consists of linear content and network similarity (the weight is 0.5 for each) and adheres to the following rules: 1) The content similarity uses the content-based cosine similarity (?, ?), which uses cosine distance between two space vectors that consist of the words of parsed data. The system computes the similarity between any pair of attendees based on the HTML content (e.g. the abstract from Google Scholar, the tweets from Twitter or the general search results). 2) This study utilizes Jaccard's coefficient(?, ?) to calculate the network similarity, which uses the intersection number of the sets of common neighbors to divide the number of the union. This similarity will show the networking relation between two persons. The system further fetches the networking information from co-authors of Google Scholar and Research Gate, or the Twitter following/follower network.

2.4 Interface

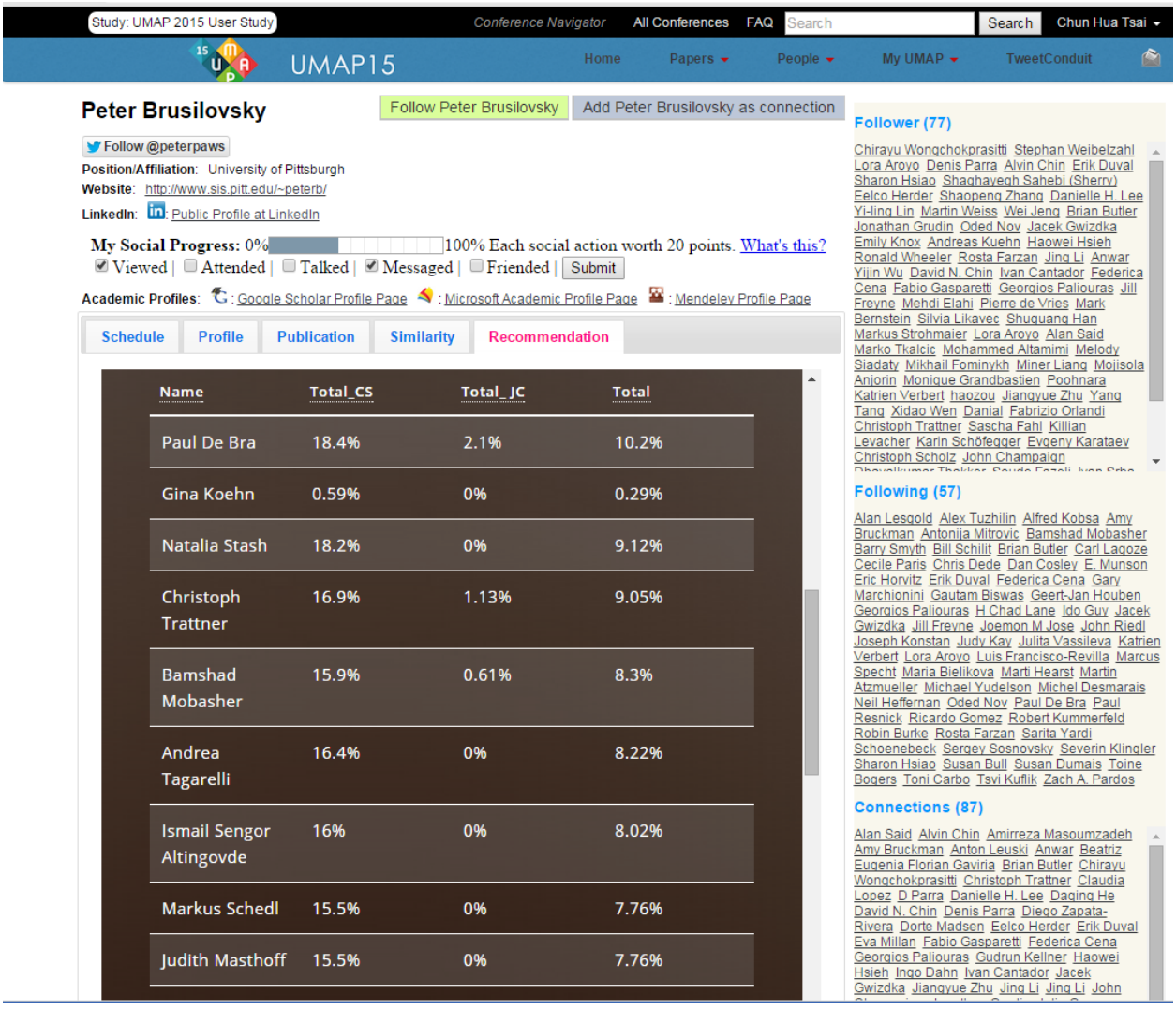

Figure 2: The screenshot of the system interface. The system is embedded in the Conference Navigator 3 system with different function tabs. The example is extracted from the UMAP2015 conference.

The system is integrated with the CN3 system as tab functions (see Figure 2). The system consists of four parts: 1) Profile Tab: this tab provides an overview of the target user's search results where the user can browse the structured data that the system used to compute the similarity and recommendation

Table 1: Survey Result of AIEDxEDM 2015 Conference

	(349 attendees, 3.4% valid response rate)				
	Not at all helpful	Not very helpful	Somewhat helpful	Very Helpful	Total Responses
Content Similarity	0 (0%)	3 (25%)	8 (66%)	1 (8%)	12 (100%)
Network Similarity	0 (0%)	3 (25%)	9 (75%)	0 (0%)	12 (100%)
Recommendation	0 (0%)	4 (33%)	7 (58%)	1 (8%)	12 (100%)
My Social Progress	1 (8%)	6 (50%)	5 (41%)	0 (0%)	12 (100%)

Table 2: Survey Result of UMAP2015 Conference

	(114 attendees, 4.3% valid response rate)				
	Not at all helpful	Not very helpful	Somewhat helpful	Very Helpful	Total Responses
Content Similarity	1 (20%)	2 (40%)	2 (40%)	0 (0%)	5 (100%)
Network Similarity	1 (20%)	1 (20%)	3 (60%)	0 (0%)	5 (100%)
Recommendation	1 (20%)	1 (20%)	3 (60%)	0 (0%)	5 (100%)
My Social Progress	1 (20%)	2 (40%)	2 (40%)	0 (0%)	5 (100%)

score; 2) Similarity Tab: this tab provides visualized word-cloud of the target user's publication and the content similarity percentage. The user can also explore the network similarities between themselves and other conference attendees through and interaction bubble (an interactive chart); 3) Recommendation Tab: this tab delivers the recommendation results. The recommended list is ordered by the total score between current user and target attendee. It also provides the sub-scores of content similarity and network similarity. It allows users to sort the list by criteria and search keywords; and 4) My Social Progresses Tool: this tool helps users update their connection progress with the target attendees. The system defines five levels of social connections of viewed, attended, talked, messaged and friended.

3 Preliminary Result

To examine the effectiveness of this system, we deployed the system at two major conferences: AIEDxEDM 2015[7] and UMAP2015[8]. The AIEDxEDM 2015 is a joint conference of AIED (Artificial Intelligence in Education) and EDM (Educational Data Mining). The attendee size is twice than the ordinary conference and the attendees' research interest background are diverse. The UMAP2015 is an international conference of User Modelling, Adaptation and Personalization held in 2015. It is a professional conference for the domain's community members and experts. The attendees are largely share a similar expertise and background.

We sent out a questionnaire to all the attendees of both conferences (see Table 1 & Table 2). The response rates are around 3.4% to 4.3%. According to the survey results, the majority of respondents indicated the system was effective. At both conferences, 40% - 75% of respondents considered the system to be somewhat helpful. The content similarity, network similarity and recommendation gathered more positive feedback than the social progress tool. Overall, the survey for AIEDxEDM 2015 received more positive feedback than the UMAP2015. In fact, only 8% of all feedback for the social progress tool at AIEDxEDM 2015 was negative, whereas 20% of feedback for all functions at UMAP2015 was negative. This may be due to the differences in attendees' backgrounds between the two conferences. In other words, this system may be more helpful to attendees of a diverse conference or group.

[7]http://perseo.lsi.uned.es/aied2015/
[8]http://umap2015.com/

4 Summary and Future Works

This study proposed a novel approach to integrate global search results into a personalized people recommendation system. We adopted the person/name/affiliation as a query and processed the search results through five different customized parsers. This approach solves the cold-start issue in recommendation systems and leverages the cross-domain information to provide a better recommendation results. We embedded this system into an existing conference navigator system as tab functions and deployed the system into two main international academic conferences. The survey results indicate a major 40% - 75% positive feedback on the system's effectiveness. Our study results also help us to understand the social interactions of an academic conference.

In future work, we plan to focus on two research topics: 1) Extending the current work to a social support system for conference new comers (e.g. Junior scholar). The system design should be personalized based on the conditions and criteria for the social needs of junior and senior members of the research community are different; 2) Building a cross-domain recommendation model based on the global search results. The data sources that can be accessed from the search engine are varied. We need to develop an approach that utilizes the accessible resources in order to estimate the remainder of unreachable information.