PhD Thesis Proposal Form
China Scholarship Council (CSC)/ENS Rennes
Call for projects 2018

FIELD open

Thesis subject title:
Social graph data mining under uncertainty using belief functions

- Laboratory name: IRISA
- PhD supervisor (contact person):
  - Name: Arnaud MARTIN
  - Position: Professor
  - E-mail: Arnaud.Martin@irisa.fr
  - Phone number: +(33)2.96.46.94.60

- Thesis proposal (max 1500 words):

The last decades have seen a surge of interest in graph data mining problems that take into consideration different forms of uncertainties. Successful existing approaches mainly rely on classical theories, such as probability theory, to deal with the uncertain information in real-world graphs. However, in recent few years, new uncertainty theories have emerged. In particular, the theory of belief functions [1] has received growing interest due to its richer representation of uncertainty and imprecision compared to probability theory as well as its higher ability to combine pieces of information.

This PhD work will focus on graph-based data mining for social network analysis. Online social networks allow only limited access to their data which generates more imprecision and uncertainty for the social network analysis fields [2,3].
The primary areas to investigate are user clustering and social recommendation. Precisely, in a first stage, we intend to group the users in the network using multiple information from complementary sources. These kinds of knowledge can be crawled from websites and is often full of uncertainty.

Then, in a second stage, we are planning to study the possibility of mixing this first line of work with recommendation of handling soft ratings and users' non-deterministic class structure.

Over a longer term, the goal is to develop a versatile and flexible framework for addressing uncertainty in social network analysis.

Detailed thesis topic

1 Active social user clustering based on uncertain information fusion

With the pervasive use of social networks, people become more relying on social media to get information and news. This also offers interesting research opportunities for analysing the behaviour and interactions between/among users. Nowadays, interactions between users are not only limited to social relations, but also to some activities, such as reading and writing. Thus, multiple and complementary information sources are available for characterising users. One task that could benefit from the integration of those multiple sources is community detection, which can also be seen as clustering problems on graph data sets.

Most graph clustering techniques have disregard the effect of information aggregation and continue to focus only the topological structure of graph. This project will focus on how to take advantage of the multiple social and content-based informationoriginated in social networks for improving the performance of the community detection algorithms. The theory of belief functions is widely used for the task of information fusion, and has already been adopted to describe the uncertain community structure in graphs [4,5]. The current evidential clustering method is of high complexity. It is of great value if an efficient evidential clustering approach can be designed for large data sets.

The content information included in the network can be attributes, user behaviors, etc, which could be crawled from the websites. Consequently, it will bring about a large amount of uncertainty and imprecision. The information from different views might be of high conflict. In this project, we will study the user clustering problems in social networks in the perspective of information fusion using the theory of belief functions.

Another problem we will consider in social user clustering is that there may be some potential prior information in social networks, for example, some pairwise constrains and some pre-labeled nodes. These knowledge can be obtained from the websites, especially, from the online crowdsourcing platforms [6]. We will study how to use a small amount of supervised information to improve the clustering accuracy as much as possible, and design a online community detection approach that can interactive with the crowdsourcing applications.

The main challenges in this topic are as follows:

- How to incorporate the multiple uncertain information in the social network in the user clustering process?
- How to design an efficient evidential clustering approach applicable for large data sets?
- How to actively select the potential supervised information to guide the community detection process?

2. Evidential recommendation incorporating trust and community information

Recommender systems are widely used to predict the ``rating'' or ``preference'' that a user would give to an item in e-commerce. Collaborative filtering (CF) is one of the most successful recommender approach which can recommend items to a target user based on the opinions of his/her like-minded neighbors. As the size of social networks is often quite large and there are millions of users, it is quite complex to find the neighbors of a target user. Clustering-based CF can be a good way to make the recommendation method more efficient. But it still suffers from relatively low accuracy and coverage [7].

The rating information provides by users are very important in recommendation systems. But it can be uncertain and imprecise, or even noisy. For example, there might be some ratings not provided by the real customers but by the friends of the sellers. Also, some users' ratings are from their own subjective view. As there are usually so many items, there are relatively few items that have been rated by one user. As a results, some (cold) users can not be targeted due to insufficient data. We will study how to express the ratings in the framework of belief functions to describe the uncertainty included in the rating and to predict the missing ratings or trust relationships from the neighbors.

This project will try to use the evidential clustering method designed in the first part of this PhD work to iteratively group users using multiple information in the network. The ratings provided by the users can be seen as one kind of content information. The evidential community structure enables us to have a deeper insight into the network data set. As a result, it could help us well perform the task of recommendation.

The main challenges in this topic are as follows:

- How to express the uncertainty in the ratings and to predict the missing information in the sparse rating matrix using the neighbourhood information?
- How to use the evidential user clustering results to guide the recommendation process?

Reference

[1] G. Shafer. A Mathematical Theory of Evidence. Princeton university press, 1976.
[2] E. Adar, C. Re. Managing Uncertainty in Social Networks. IEEE Data (base) Engineering Bulletin, 2007, 30:15–22.
[3] S. Jendoubi, A. Martin, L. Liétard, et al. Two Evidential Data Based Models for Influence Maximization in Twitter. Knowledge-Based Systems, 2017, 121:58–70.
[4] K. Zhou, A. Martin, Q. Pan, et al. Median Evidential C-means Algorithm and its Application to Community Detection. Knowledge Based Systems, 2015, 74:69–88.
[5] K. Zhou, A. Martin, Q. Pan, et al. SELP: Semi-supervised Evidential Label Propagation Algorithm for Graph Data Clustering. International Journal of Approximate Reasoning, 2018, 92:139–154.
[6] Y. Yan, R. Rosales, G. Fung, et al. Active Learning from Crowds. International Conference on Machine Learning, ICML 2011, Bellevue, Washington, Usa, June 28 - July. 2011:1161–1168.

[7] G. Guo, J. Zhang, N. Yorke-Smith. Leveraging Multiviews of Trust and Similarity to Enhance Clustering-based Recommender Systems. Knowledge-Based Systems, 2015, 74:14–27.

- Publications of the laboratory in the field (max 5):

Kuang Zhou, Arnaud Martin, Quan Pan, Zhunga Liu, SELP: Semi-supervised evidential label propagation algorithm for graph data clustering, International Journal of Approximate Reasoning, 92, pp.139-154 (2018)

Zhun-Ga Liu, Quan Pan, Jean Dezert, Arnaud Martin, Combination of classifiers with optimal weight based on evidential reasoning, IEEE Transactions on Fuzzy Systems, 2017.

Siwar Jendoubi, Arnaud Martin, Ludovic Liétard, Hend Ben Hadji, Boutheina Ben Yaghlane, Two evidential data based models for influence maximization in Twitter. Knowl.-Based Syst. 121: 58-70 (2017)

Kuang Zhou, Arnaud Martin, Quan Pan, Zhun-ga Liu, ECMdd: Evidential c-medoids clustering with multiple prototypes. Pattern Recognition 60: 239-257 (2016)

Kuang Zhou, Arnaud Martin, Quan Pan, Zhunga Liu, Median evidential c-means algorithm and its application to community detection. Knowl.-Based Syst. 74: 69-88 (2015)

- Joint Phd (cotutelle) : YES
- Co-directed PhD : YES

In case of a co-directed or a joint PhD, please detail:

- Partner university name: Northwestern Polytechnical University, Xi'an, China
  - Laboratory name and web site: Key laboratory of information fusion technique, http://zdhxy.nwpu.edu.cn/
- PhD co-director (contact person):
  - Name: Zhunga LIU
  - Position: Professor
  - E-mail: liuzhunga@nwpu.edu.cn
  - Phone number: 0086-029-88431371

- Provisional duration and timetable of the PhD student's stay at ENS Rennes:

  We propose 2 years in France por the PhD student and the rest of the time in China.

- If previous collaborations with the Chinese co-director/university, please detail:

  The collaboration between IRISA lab and the Northwestern Polytechnical University began in 2013 with the co-supervision of Kuang Zhou. His thesis, defended in Rennes in July 2016 allows the publication of 5 journal papers in the best international journal of the domain and 8 communications in international conferences. Moreover, we develop a package available on the CRAN website.
  We also organized, last summer, a school in Xi'an on the theory of belief functions.

- Interest of the Joint PhD for the French co-director, for his/her laboratory, for ENS Rennes:

To develop the collaborations between IRISA and the Northwestern Polytechnical University with this joint PhD will for me the way to continue the very good developments (theoretical and practical) around data mining on social network with uncertainty theories. This subject is still very challenging and is one of the main topic of the DRUID team.
For IRISA, this collaboration can place the laboratory on the first international rank on this very important subject of the data mining on social network.
This collaboration can be a good opportunity for ENS Rennes to welcome best Chinese students from Xi'an and to develop research collaboration in other fields.

Date: 2018/01/12

Signature of the PhD director

Name and signature of the Laboratory director
Jean-Marc Jézéquel, Director