Knowledge-Based Content Suggestions on the Social Web

EDITE 2012 PhD Proposal - Summary Sheet

Theme: Content, Knowledge, Interaction

PhD Advisors: Bogdan Cautis and Pierre Senellart are both associate professors in computer science at Télécom ParisTech. Both supervisors work in the area of Web data management: database theory, Web mining and social network analysis. Their full CVs are joined.

PhD Candidate: Georges Gouriten has been hired as a R&D engineer at Télécom ParisTech in July 2011. He holds a Master's degree in engineering from Mines ParisTech and has a record of international experiences. At Télécom ParisTech, he has acquired a research experience that led to a first publication in a major international conference (WWW, developer track). His full CV is attached.

Doctoral School: This PhD will be attached to Télécom ParisTech and EDITE de Paris. The funding will be in the form of a *contrat doctoral* CDD, possibly complemented with teaching activities.

Host Research Team: Télécom ParisTech, INFRES department, IC2 group, DBWeb team `http://dbweb.enst.fr/`

Keywords: Social Web, Semantic Web, Knowledge Base, Linked-based ranking, Recommendation, Community Detection

Abstract

Billions of users daily use Web-based social networks. On those platforms, people share thoughts, opinions, documents, or events. In the process, large amounts of data are created. An important part of this social data on the Web is accessible. We have the opportunity to use this data to help us expand our knowledge on individual or global questions. Imagine *Stevie* is part of a social network and fan of jazz pianists, given his social context and preferences, who should be the next person he contacts?

This vast amount of human-generated data has the potential to open the way to many advances in artificial intelligence. However, it is still very difficult to process intelligently and at large-scale. Most state-of-the-art techniques do not take into account the subtlety of the social context or rely on keyword-based information retrieval. There has also been important progress in representing human knowledge in a way that is understandable by machines. We propose to create systems able to access data from the social Web and make intelligent content suggestions given a user, an entity from a knowledge base, or both. It requires us to work on ambitious challenges in many research areas, in particular *linked-based ranking*, *community detection*, and *recommendation*. Moreover, scalability, data acquisition, and data heterogeneity are, in our context, important issues.

EDITE 2012 PhD Proposal - Detailed Description

PhD Advisors: Pierre Senellart (MCF, Télécom ParisTech) and Bogdan Cautis (MCF, Télécom ParisTech). Fully joint supervision.

PhD Candidate: Georges Gouriten (currently R&D engineer at Télécom ParisTech).

Theme: Content, Knowledge, Interaction

Title: Knowledge-Based Content Suggestions on the Social Web

1 Scientific Background

The creation of Facebook in 2004 started the spectacular rise of Web-based social networks, now daily used by billions of users. On those platforms, people share thoughts, opinions, documents, or events. In the process, large amounts of data are created. Twitter, for instance, claimed an average of 140 million tweets sent per day in March 2011.[1]

An important part of this social data on the Web is accessible, either because it is public, or because access can easily and securely be granted by a user. This data can help us guide our thinking on many individual or global questions. Imagine *Stevie* is part of a social network and fan of jazz pianists, given his social context and preferences, who should be the next person he contacts? Now, suppose *Stevie* wants to make an article on Keith Jarrett, who is the globally most relevant person he should write to for an interview?

The potential of this data opens the way to major progresses in artificial intelligence. However, it is still hard to process intelligently, especially at large-scale. Most approaches do not take into account the subtlety of a social context, for instance disregarding the fact that the strength of a friendship link can be variable. It is also common to use simple keyword-based information retrieval, thus ignoring that jazz and bebop are very related concepts.

In the meantime, there has been important progress in representing human knowledge, under the form of standardized knowledge bases, understandable by machines. Those knowledge bases now contain large amounts of human knowledge, represented as sets of statements – such as "Keith Jarrett is born in Allentown" – linking identified entities – "Keith Jarrett" and "Allentown".

We want to build systems able to access data from the social Web, and make intelligent content suggestions relatively to a user, an entity, or both. It requires us to work on ambitious challenges in many research areas, in particular *linked-based ranking*, *community detection*, and *recommendation*. Moreover, scalability, data acquisition, and data heterogeneity are, in our context, important issues.

Our proposal is organized as follows. We continue to describe our scientific context and define precisely our representation for social data and knowledge bases. Then we detail our research challenges, and the related work. Finally, we explain why our proposal is in conformity with selection criteria.

Social data and knowledge bases

We want to combine social data from the Web and human knowledge, formalized in knowledge bases. In this section, we describe precisely our data representation.

[1] Twitter (2011), "#numbers", `http://blog.twitter.com/2011/03/numbers.html` [accessed April 25, 2012]

Social data on the Web Social data is data emitted by a user. In our model, we consider sets of users. A user can be a person or a social entity. A social entity is a group that cannot be reduced to the sum of its individual members, for instance a company. Most social platforms allow to define links between users, such as the friendship links on Facebook. Those can be directed (X follows Y), or undirected (X and Y are friends). It is thus common to model a social network as a graph where nodes represents users while edges represents links between users.

Additionally, linked to each user are a profile – the user's description – and a set of activities. An activity is data created at a specific time by the user, such as a tweet. This temporal division is important to be able to split data into items of coherent content. Though our focus will be on text and semi-structured data, an activity or a profile can also contain visual or sound elements. An activity can refer to other activities (e.g., a retweet), or to other users (e.g., a name mention on Facebook). We can therefore extend our model with new types of nodes, the activities, and directed edges, activity–activity or activity–user.

The World-Wide Web currently offers the most dynamic and comprehensive source of social data. We plan to experiment content suggestion with different sources such as social networking websites (e.g., Twitter, Google Plus, or Delicious), blogs, forums, or newspaper websites allowing comments (e.g., lemonde.fr).

Knowledge bases Knowledge bases are formal representation of generic knowledge. For computer scientists, they often consist in sets of entity–relation–entity triples, such as { "Herbie Hancock", isA, "Jazz pianist" } or { "Herbie Hancock", isBornIn, "Chicago" }. Some classes of entities can be defined as well as internal logics. For instance, the domain of isBornIn is the class People, and its range is the class City. We plan to use large state-of-the-art knowledge bases, such as YAGO [17]. A knowledge base can be modeled as a graph of labeled nodes, the entities, and labeled edges, the relations.

Connecting the two We can connect user profiles or activities to one or several knowledge bases thanks to semantic annotation. Advanced techniques achieve near-human performances [15]. For instance, from an activity such as { user_u, date_d, "Herbie Hancock is my favorite jazz pianist" }, we can infer that *Herbie Hancock* is the well-known jazz pianist. It is therefore possible to link this activity to the entity *Herbie Hancock* of an ontology.

Overall, our data representation is a large-scale heterogeneous graph with directed labeled edges. It is large-scale since there are many users, activities, entities, and many edges between them. To give an order of magnitude, YAGO contains more than 100 million relations and we could potentially get hundreds of millions of activities, for example tweets, belonging to millions of users. It is also heterogeneous as we combine different kinds. In addition, we could imagine adding other layers, for instance to consider sentiment analysis.

2 Research Challenges

There are many exciting research opportunities in integrating social data with knowledge bases. Let us take a concrete example. *Maurice*, a user, is very much interested in the entity *Herbie Hancock*. Our aim is to suggest him some more content he would like. We could start looking for users related to him and interested in topics close to *Herbie Hancock*. This implies having a good user-centric ranking that would combine an idea of social proximity and interestingness relatively to an entity. Then, we want to make relevant recommendations among those users and pick different possibilities from different social or topical context.

Ranking, recommendation, and clustering are fairly mature research areas, but data heterogeneity, data size, the use of knowledge bases and a clever way to combine all these techniques make it possible to innovate on many points and achieve unattained pertinence of results.

Given the amount of data we expect to deal with, it is also important to develop scalable systems. To this end, we shall focus on efficient sequential algorithms, distributed algorithms, intelligent indexing, and approximation techniques.

Multi-dimensional link-based ranking In our example, we want to rank users close to *Maurice* and interested in *Herbie Hancock*. For each user, we could give a score that combines those two dimensions. Proximity to *Maurice* could be expressed as the inverse of the length – or some other function – of the shortest path between the user and *Maurice* in the social graph.

Interest of a user X relatively to *Herbie Hancock* could be an aggregation of measures for each activity related to X. An activity could be indirectly related to a user, through a mention for instance, and we could use a factor of proximity. To tell how interesting an activity is relatively to an entity, we could consider taking into account popularity – higher if it has been shared by others – and closeness to the entity – higher if the text contains the entity, or related topics.

This is a first example of ranking among many others. Who are the most interesting users relatively to *Herbie Hancock*? What are the most interesting entities for *Maurice*? What are the most interesting activities related to *Herbie Hancock*? With this very rich setting, we hope to find new ranking algorithms that we plan to evaluate with rigorous user-centric and performance tests. We would also probably have to work on distribution and approximation techniques, that would allow us to rank large datasets efficiently.

Social and knowledge-based recommendations With personalized data, we can learn about user preferences and think on how to make interesting recommendations. In our representation, we can propose recommendations at several levels. We could suggest to link to another user, given preferences and social proximity, as well as to recommend an artist, because the activities and the others users he is linked to are very related.

We want to diversify recommendations so that not only generic top items are suggested but top items from different clusters. In our example, *Maurice* appears to be very interested in jazz pianists. Instead of promoting the three most-popular jazz pianists, it could be more interesting to suggest a popular jazz pianist from *fusion jazz*, another from *bebop* and another from *cool jazz*. Similarly, instead of recommending him to be in contact with members of the *International jazz* association, he will also be proposed to get in touch with members of the *Jazz Paris club*, the *Jazz afficionados confederation* and the *Herbie Hancock fan club*.

Community detection Detecting communities is important to make more accurate estimations, diversify results or distribute computation. We think of two main approaches. On one hand, we could to look at social communities, based on social interactions. On the other hand, we could identify topical communities from common interests in similar subjects. We would then be very interested in studying how the two types of clustering can interact. We will also focus on doing community detection in a scalable manner and oriented towards practical applications.

Results explanation In our scenario, analysis can become complex. Therefore, we could think on how to explain results to users. Back to our example, we would want to explain to *Maurice* that *Jaco* has been selected because he is socially close and very-interested in *Herbie Hancock*. This could be mean highlighting a relevant sample of common contacts, showing *Jaco*'s two popular

activities explicitly mentioning *Herbie Hancock*, and saying that eighty percent of its activities are related to fusion jazz.

Data acquisition Our data sources are typically Web services, they are accessible via specific HTTP requests. We want to be able to fetch interesting datasets efficiently, this implies solving several problems. Web services being very diverse, how is it possible to conveniently interact with them? It seems promising to extend our work on the API Blender [5], a system that allows to interact with multiple Web services. Given the constraints imposed by many Web services (e.g., limited rate), how can we have performing fetching strategies? We plan to develop information retrieval techniques adapted to social Web services with strong constraints. Besides, a first intuitive approach is to fetch the data and then analyze it offline, can we imagine online strategies?

Data shaping We fetch raw data from the *wild* Web, and we want to be able to shape it in a convenient manner. The data format can be, for instance, HTML, JSON, or XML, and with different schemes. How can we integrate everything? We plan to use manually and possibly automatically generated lightweight schema description formats. Another issue is, once the data is integrated, how to efficiently run semantic annotation? We plan to rely on state-of-the-art tools such as GATE [2].

Data cleaning Data can sometimes be redundant or implicitly related. For instance, how to detect that user x on Twitter is the same as user y on Facebook? We can think of near-duplicate detection approaches that take into account the different dimensions of our data: for instance, we can compare activities on different platforms and, if they match, it would then probably be the same user on both platforms.

There is also some spam that we will want to filter. The social context should be helpful in this effort, spam being by definition not interesting for most people.

3 Related Work

Linked data analysis Most of our efforts will consists in analyzing data that can be represented as a set of connected items, this domain has inspired much research. As early as in 1953, Katz modeled a vote network as a matrix [7]. In 1972, Garfield made a first analysis of journal citations [3]. Many other important contributions in this field followed, they are well described in a survey by Kleinberg in 1999 [8] and more recently by Getoor and Diehl [4].

Ranking Ranking consists in ordering data items depending on their relevance regarding a given context. The persistent popular appeal of Web search engines highlights the importance of data ranking. A major contribution to this large domain has been the PageRank algorithm [14]. It has been designed for a directed graph model and has proved very successful in identifying relevant Web pages when applied to the Web graph. PageRank has led to many related research and many variants adapted to different situations, among them a topical version [6] and its generic formulation [13].

Our situation is more complex than state-of-the-art Web ranking techniques. We need to be able to rank while taking into account several interlinked and heterogeneous dimensions. An interesting effort to add a social dimension to standard information retrieval techniques was made

recently [16]. It covers some related aspects combining a social and a keyword score but is limited to these elements.

Recommendation Research on recommender systems has been intensely driven by their application on commercial platforms and is rich of many techniques. For a long time, there has been two main approaches, collaborative filtering where users' behaviors are compared, and content-based filtering where items similar to the ones liked by the user are found. Recently, hybrid approaches have been successfully deployed. A comprehensive survey was published in 2005 [1].

Some of these techniques have also recently been successfully parallelized [9], but to the best of our knowledge, none of these approaches consider a multi-dimensional setting such as ours.

Community detection A seminal work by Newman [12] in 2006, drew a parallel between clustering in networks and matrix calculus with the principle that users belonging to the same community are densely connected while users belonging to different communities are not. Since then, there have been several works applying clustering techniques to social networks [11] or proposing parallel approaches [10].

4 Conformity with Selection Criteria

The DBWeb team in general, and the two PhD advisors in particular, have strong background in data management and social data analysis. We offer concrete research directions partially inspired by fundamental research in computer science, that could have large impact in several fields of computer science.

At the international level, the PhD research would be carried out in collaboration with the following research institutes:

Max-Planck-Institut für Informatik, Saarbrücken through existing collaborations, information extraction and construction, as well as maintenance, of real-world knowledge bases (YAGO, MENTA).

The University of Hong Kong through the PROCORE France–Hong Kong project on *Private Recommandation Systems*, joint between Pierre Senellart and T.-H. Hubert Chan.

Rutgers University through an existing collaboration with Amélie Marian on the construction and querying of knowledge bases of personal information.

Both Bogdan Cautis and Pierre Senellart have supervised several PhD candidates on various aspects of Web data management and will have their *Habilitation à diriger les recherches* at the time of this project. Georges Gouriten has demonstrated his ability at carrying research work successfully in the field of social Web data management as a R&D engineer in the DBWeb group.

The ambitious goals of our PhD proposal perfectly match the spirit of the EDITE call, which is about basic, foundational, and long-term research, yet applied to real-Web data. One potential outcome of a three-year PhD could be the creation of a start-up that would apply the techniques developed during the course of the PhD to propose state-of-the-art social data mining or querying interfaces to the social Web.

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6), 2005.

[2] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. Evolving GATE to meet new challenges in language engineering. *Nat. Lang. Eng.*, 10(3-4), 2004.

[3] E. Garfield. Citation analysis as a tool in journal evaluation can be ranked by frequency and impact of citations for science policy studies. *SCIENCE*, 178(4060), 1972.

[4] L. Getoor and C. P. Diehl. Link mining: a survey. *SIGKDD Explor. Newsl.*, 7(2), 2005.

[5] G. Gouriten and P. Senellart. API Blender: A uniform interface to social platform APIs. In *Proc. WWW*, 2012.

[6] T. Haveliwala. Topic-sensitive PageRank: a context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4), 2003.

[7] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18, 1953.

[8] J. M. Kleinberg. Hubs, authorities, and communities. *ACM Comput. Surv.*, 31(4es), 1999.

[9] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

[10] K. Macropol and A. Singh. Scalable discovery of best clusters on large graphs. *Proc. VLDB Endow.*, 3(1-2), 2010.

[11] N. Mishra, R. Schreiber, I. Stanton, and R. E. Tarjan. Clustering social networks. In *Proc. WAW*, 2007.

[12] M. E. J. Newman. Modularity and community structure in networks. *Proc. National Academy of Sciences*, 103(23), 2006.

[13] Y. Ollivier and P. Senellart. Finding related pages using Green measures: An illustration with Wikipedia. In *Proc. AAAI*, 2007.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.

[15] L. Reeve and H. Han. Survey of semantic annotation platforms. In *Proc. SAC*, 2005.

[16] R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, and G. Weikum. Efficient top-k querying over social-tagging networks. In *Proc. SIGIR*, 2008.

[17] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proc. WWW*, 2007.