Description des travaux

Abstract: The project aims at models, algorithms and platforms for Web data featuring: *structured documents*, *semantics* and *social* context in which data is produced and exploited. A starting point is the XR model developed by Inria and Paris Sud for structured documents with semantic annotations, to which Telecom brings its expertise on social-aware ranked search.

PhD topic (max 3 lines) : Social-aware top-k keyword search over rich Web content

CV of PhD supervisors

The three co-supervisors are B. Cautis (Telecom ParisTech), F. Goasdoué (U. Paris-Sud) and I. Manolescu (Inria Saclay). Their CVs have been provided earlier in this document.

Doctorant :
Y a-t-il un doctorant connu ? **non**

Detailed PhD topic description

Social-aware top-k keyword search over rich content

Context: annotated documents. Tools for authoring electronic content and sharing it through the Internet are very widely adopted by now. First blogs, and then social networks, grew more or less in parallel with the major media providers' move towards allowing users to record their opinions next to the articles. These technical means to hold back-and-forth conversations, as well as the Linked Open Data movement[1], in which public and private institutions publish their data for transparency and accountability reasons, have sparked new usages of the Web and in particular of Web data. As another example, public figures rely heavily on social networks to communicate their positions to the public, collect feedback, survey opinion trends, etc. At the same time, individuals interested in a particular topic or action (e.g., the economic impact of a certain policy) can comb the Web for bits of information, connect, interpret, annotate and re-share them. For instance, such data gathering and fact checking have come at the core of "data journalism", pioneered, e.g., in Europe by The Guardian[2] and growing through efforts such as FactCheck[3], Politifact[4], and similar French sites[5].

The increased interest of the public at large for reading, analyzing, commenting, and crossing information raises the need for (i) integrated, generic and open tools to gather and share content and annotations, and (ii) effective and scalable tools for querying such data. Wide-audience tools must be *integrated* to spare the user the effort of gluing several software components. They need to be *generic* to handle many types of input. For instance, if one wants to analyze some tweet stream, one very likely needs to archive and analyze Web pages, RSS feeds or domain ontologies. They should also be *open* w.r.t. the supported data formats, so that they are not only information sinks but also information sources, and w.r.t. the architecture, so that they can be easily customized and extended. *Effectiveness* in retrieving the most relevant content is of foremost importance in order to help users make sense of the data, and to provide them well-defined, customized views over it. Finally, *performance* is the recurrent aspect that must be guaranteed at the scale of Web applications, over large repositories of annotated data and complex communities of users.

[1] http://linkeddata.org
[2] http://guardian.co.uk/data
[3] http://www.factcheck.org
[4] http://www.politifact.org
[5] http://www.liberation.fr/desintox, http://decodeurs.blog.lemonde.fr

The XR model. To address the need for rich content management, within the Oak team, we have proposed XR, a language and model based on the W3C standards XML for representing structured Web documents, RDF for encoding facts and more generally Semantic Web data, and RDF Schema for encoding knowledge (i.e., ontologies) [5], under the single paradigm of *annotated documents*. In [5, 6] we have defined the language and its associated query language, XRQ. One follow-up currently under minor revision for an international journal analyzes strategies for the efficient evaluation of XR queries. A demonstration on how the XR model and platform can be applied to warehouse and analyze controversies on the Web will be shortly presented at the ACM SIGMOD conference [4]. More details about our project can be found at http://tripleo.saclay.inria.fr/xr.

Goals. The PhD topic we propose considers the problem of *social-aware top-k keyword search over XR documents*. As more and more users have the opportunity to create, publish, share, and interact upon rich content, novel query models and query processing techniques are necessary in order to best answer the users' informational needs. When data and annotations have a social nature – being the result of collective contributions by individuals and communities – in the presence of variate social relationships, query answering should be done in a *network-aware* manner. This means that answers should be relevant both textually and with respect to measures of social proximity and similarity. Intuitively, the interests or viewpoints of my friends (closest users) are correlated to mine; therefore, relevance models that assess the correspondence between a data item and a query should include the social attributes of that data.

We will associate with the notion of social network a generic interpretation, as a user graph whose edges (links) are labeled by *social scores*, which can capture any appropriate measure of the proximity or similarity between two users. This means that, even for applications where an explicit social network does not exist or is not exploitable, we may use user actions and interactions to build a network based on certain metrics that can, for instance, be indicative of affinity and similarity.

As each data item (e.g., a tag, an annotation, etc) originates at one social entity in the network, early-termination search techniques must jointly explore the social and data space; this is inherently difficult. In addition, as users act as both procedures and consumers of information, it is highly desirable to support techniques by which users are given complete control (full personalization) over the model choices that determine how results are influenced by social links. Indeed, recent studies on result relevance with respect to information needs in social applications show that a one-size-fits-all query model is often unsatisfactory. Furthermore, social and collaborative data is of a highly dynamic nature. This is of course due, on one hand, to the dynamics of user-generated content. On the other hand, it is also due to the fact that we are dealing with highly interconnected data; this means that one change over a single data item may have an impact in a much broader scope.

Due to these aforementioned aspects, the problem we study is a complex one. While efficient techniques for query processing over rich data, combining semi-structured and semantic content, can be built based on existing approaches, these are not useful when the social dimension becomes a central focus. Intuitively, we cannot build an index for each user or social perspective that may have to be answered. This is is why, besides exact (complete) techniques, which in certain applications scenarios may prove to be too slow for practical purposes, we also intend to consider techniques that are only approximate, trading precision for efficiency in an acceptable way.

Main related work. While recent literature has proposed techniques for top-k keyword search over textual documents [3] and semi-structured documents [2], as well as over data from

social applications [1, 8, 12] (e.g., in social bookmarking[6] or micro-blogging[7]), this problem remains largely unexplored for rich content – in the form of semi-structured data with semantic annotations – which is produced and consumed by members of a social community.

Expected results. The goal of this thesis is to develop effective query models and efficient algorithms for content retrieval, leveraging the three key dimensions of the data: semi-structured / XML format, semantic annotations, and social authorship and social relations. This will be achieved through

1. A formal study on query models, semantics and score functions for XR data having social features,
2. The study of sound and complete algorithms for information retrieval based on the proposed query semantics and score models,
3. The study of scalable approximate algorithms for information retrieval based on the proposed query semantics and score models,
4. The integration of the proposed techniques into a prototype content-retrieval engine, and evaluation of their effectiveness on real-world data at Web scale.

Therefore, the project will have two main facets: (i) a formal, foundational study of query models and optimality for information retrieval in large-scale web applications, and (ii) on the practical side, the implementation and deployment of the resulting algorithms for content retrieval.

Success criteria. Our contributions in this PhD research project must lead to query models that are rich and flexible enough to accommodate user and application requirements in many contexts, to techniques that are efficient, scalable, with high precision levels (in the case of approximate results), as well as to query results that are relevant and trustworthy.

PhD candidates. Interesting candidates should have good background in the areas of databases, information retrieval and knowledge representation. Familiarity with Web standards (XML, RDF) is a plus, but these can be learned on site. Demonstrable good academic results in topics such as logics, artificial intelligence and advanced algorithms are also strong arguments in favor of a candidate.

Partnership. The project brings together complementary expertise in the topics involved. The basis of the XR language has been laid out by the Inria and Paris Sud partners during the second half of the PhD thesis of Julien Leblay, co-supervised by F. Goasdoué and I. Manolescu and scheduled to defend at the end of 2013.

The competence on social and top-k search is brought by the Telecom ParisTech partner (B. Cautis) having worked in this area notably as part of the defended PhD thesis of S. Maniu.

The Oak team to which I. Manolescu and F. Goasdoué participate is also a partner of a joint Inria International research team named OakSad starting in 2013 with the database group from UCSD, notably Alin Deutsch. B. Cautis is also involved in the proposal as a collaborator, as he has worked with Alin Deutsch in the past and collaborations are still ongoing. More broadly, we all have common background in semistructured database optimizations and notably on materialized view-based query evaluation, and follow-up work on XR is part of the three-years OakSad project. The Inria international team only finances visits among partners (15.000 Euros for 2013).

Relevance to Digiteo and Digicosme.

[6] http://www.delicious.com
[7] http://www.twitter.com

Digiteo The project pertains to the data and information management area and as such is related to Information Systems, part of the *Programming, Software Engineering and Information Systems* theme of Digiteo.

Digicosme The project is central to our activity within the DigiCosme Labex, to which Inria Saclay, Paris Sud and Telecom ParisTech are key partners, notably through the involvement of the Oak Inria team, Databases and IASI teams of LRI/Paris Sud, and DBWeb Telecom ParisTech teams to the **DataSense** axis. Within DataSense, I. Manolescu animates the Task 1: **Scalable, expressive and secure tools for large-scale data**. The three teams listed above also participate to Task 2: **Making sense of complex, heterogeneous data**.

The topic described in this project is related to both DataSense tasks, through its interest in large-scale data management and expressive models for Web data.

Bibliography

[1] S. Amer-Yahia, M. Benedikt, L. V. S. Lakshmanan, and J. Stoyanovich. Efficient network aware search in collaborative tagging sites. *PVLDB*, 1(1):710–721, 2008.

[2] L. J. Chen and Y. Papakonstantinou. Supporting top-K keyword search in XML databases. In *ICDE*, pages 689–700, 2010.

[3] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *J. Comput. Syst. Sci.*, 66(4):614–656, 2003.

[4] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, and I. Manolescu. Fact-checking and analyzing the web (demonstration). In *SIGMOD*, 2013.

[5] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, and S. Zampetakis. Growing Triples on Trees: an XML-RDF Hybrid Model for Annotated Documents. In M. Brambilla, F. Casati, and S. Ceri, editors, *First International Workshop on Searching and Integrating New Web Data Sources*, Seattle, United States, Sept. 2011.

[6] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, and S. Zampetakis. Growing Triples on Trees: an XML-RDF Hybrid Model for Annotated Documents. In *Journées de Bases de Données Avancées*, Rabat, Morocco, Oct. 2011.

[7] A. Katsifodimos, I. Manolescu, and V. Vassalos. Materialized view selection for XQuery. pages 565–576, 2012.

[8] S. Maniu and B. Cautis. Efficient top-k retrieval in online social tagging networks. In *Journées de Bases de Données Avancées*, Rabat, Morocco, Oct. 2011.

[9] S. Maniu and B. Cautis. Context-aware top-k processing using views. In *Journées de Bases de Données Avancées*, Oct. 2012.

[10] S. Maniu and B. Cautis. Taagle: efficient, personalized search in collaborative tagging networks. In *SIGMOD Conference*, pages 661–664, 2012.

[11] I. Manolescu, K. Karanasos, V. Vassalos, and S. Zoupanos. Efficient XQuery Rewriting using Multiple Views. In *ICDE*, 2011.

[12] 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 *SIGIR*, pages 523–530, 2008.