Challenges in Enabling Social Application at Scale

CloudDB'12 Invited-Keynote Talk Abstract

Ashwin Machanavajjhala
Duke University
Durham, NC, USA
ashwin@cs.duke.edu

ABSTRACT

Internet users spend billions of minutes per month on social networking sites like Facebook, LinkedIn and Twitter. Not only do they create tons of data everyday in the form of posts, tweets and photos, the connections between users have given rise to new applications for social discovery and engagement. In this talk I will highlight the novel data management and privacy challenges in three such applications: (i) Feed Following, or the problem of delivering highly personalized feeds based on content generated by one's friends, (ii) Social Coordination, or the problem of jointly planning and coordinating on a task, and (iii) Social Recommendations, or recommending objects and people based on one's social connections.

Categories and Subject Descriptors

H.2.4 [**Database Management**]: Systems; H.3.4 [**Information Storage and Retrieval**]: Systems and Software

General Terms

algorithms, security

Keywords

feed following, social coordination, enmeshed queries, social recommendations, differential privacy

1. INTRODUCTION

Internet users spend billions of minutes per month on social networking sites like Facebook, LinkedIn and Twitter. Social networks have revolutionized the way we generate information online; billions of posts, tweets and photos are shared everyday. The connections between users have given rise to new paradigms for information consumption and coordination. At the same, since much of the user activity and the information shared on a social network could potentially be sensitive to individuals, data privacy is an important concern. In this talk I will call out three paradigms in particular, which present novel data management and privacy challenges – *feed following*, *personalized social recommendations*, and *online social coordination.*

Feed Following: Social networks let users *follow* a larger number of other users/entities, and then retrieve personalized *feeds* consisting of events produced by (or associated with the activities of) these users/entities. Examples include the *News Feed* on Facebook, and Yahoo! Pulse, which aggregates social feeds from Yahoo! and other sites. These follows networks have hundreds of millions of edges and billions of connections.

The basic primitive is a *feed query(u)*: return a subset of the most personally relevant events generated by users that u follows, with a premium on recency. Feed queries can either be launched explicitly, or implicitly generated due to various user actions (e.g., user login). Feed delivery is by no means a simple problem due to the combination of dense connection networks, low-latency requirements (tens of milliseconds), and extremely high event and query rates. The high skew in fan-ins (for event producers), fan-outs (for consumers), and the fact that feed queries need not be continuous, but rather are sporadic, further complicates the optimization problem.

Social Recommendations: Recommendation allow serendipitous information discovery. Social networks allow new forms of recommendations – recommendations that rely on one's social connections in order to make personalized recommendations of ads, content, products, and people. Recommendations based on social connections are especially crucial for engaging users who have seen very few movies, bought only a couple of products, or never clicked on ads. While traditional recommender systems default to generic recommendations, a social-network aware system can provide recommendations based on active friends. However, improved social recommendations come at a cost – they can potentially lead to a privacy breach by revealing sensitive information. For instance, if you only have one friend, a social recommendation algorithm that recommends to you only the products that your friends buy, would reveal the entire shopping history of that friend - information that he probably did not mean to share.

Social Coordination: While social networks like Facebook are widely used for interacting with friends, they seem less useful for planning and coordinating with other people. But, this is a significant part of what being social entails – we go to birthday parties, we play sports, we fraternize with like-minded people on even obscure topics such as fly-fishing. Further, the default method of coordinating with people by phone/email is tedious, and sometimes hard to drive to closure. Moreover, many such coordination tasks must be fluid, where users may not specify explicitly with whom

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
ACM 978-1-4503-1156-4/12/10.

they would like to coordinate. Fluid coordination appears in a number of real world applications. Consider multiplayer online gaming where users wish to form groups whose sizes can depend on the game. The user is indifferent to player identities except that they be close by (to reduce latency), and have similar Internet access speeds and game ratings. A system must be able to allow a user, of say Xbox LIVE, to specify parameters in these three key attributes and a group size (say 4 to play Halo), and match such users so that they can coordinate on the game. As a second example, consider finding 3 partners among the 2-hop neighborhood in your social network for playing doubles tennis in Sunnyvale at 4 pm on the weekend. Unlike in feed following, where hundreds of thousands of queries must be answered, in the social coordination problem, one needs to collaboratively find many groups of such queries that can coordinate. Moreover, coordination partners for a coordination query may not even be in the system, when it is first posed. These characteristics make this a novel data management problem.

2. DATA MANAGEMENT CHALLENGES

All three problems can be phrased as online optimization problems, which pose novel data management challenges due to their scale, and the correlations between different user queries. For instance, while one could try to adapt solutions based on pub/sub, caching, and materialized views for feed following and recommendation, none of these existing approaches fully exploit the unique characteristics of feed following [3]. In this talk, we will motivate and highlight three interesting research challenges – *view selection*, or the problem of choosing which partial views to materialize for efficient answering of feed queries, *view scheduling*, or the problem of when to materialize the partial view to optimize for the sporadic nature of feed queries, and *view placement*, or the problem of choosing where to place partial views in a geographically distributed system.

Again, while specific forms of social coordination appear in tools such as Meetup and in game platforms such as XBox LIVE, we introduce a more general model using what we call enmeshed queries. An enmeshed query allows users to declaratively specify an intent to coordinate with other users (who they may not know a priori) by specifying constraints on who/what/when as well as on the composition of the group, such as the desired group size. The database returns a group of users who have registered queries with matching intents. Enmeshed queries are continuous, but new queries (and not data) answer older queries; the group constraints and the ability to coordinate with unknown partners make enmeshed queries differ from entangled queries, publish-subscribe systems, dating services and nested transactions. We will present some initial work on enmeshed queries [1], and highlight the scalability challenges in implementing coordination using enmeshed queries.

3. PRIVACY RESEARCH CHALLENGES

Privacy issues become important in any social application, and they can impact either accuracy, efficiency or both. In feed following and social coordination, access control rules are essential in controlling to whom events are disseminated and who can coordinate, respectively. However, allowing fine-grained access controls result in scalability challenges as opportunities for indexing, caching and view materialization reduce. Social recommendations that are more personalized to ones social neighborhood are more accurate, but also have the potential to leak more private information about one's friends. We will present some initial work on identifying privacy utility tradeoffs in social recommender systems [2], and outline research directions towards building privacy aware recommenders.

4. REFERENCES

[1] J. Chen, A. Machanavajjhala, and G. Varghese. Scalable social coordination using enmeshed queries. *CoRR*, abs/1205.0435, 2012.

[2] A. Machanavajjhala, A. Korolova, and A. D. Sarma. Personalized social recommendations - accurate or private? In *Proceedings of Very Large Data Bases (PVLDB)*, volume 4, pages 440–450, 2011.

[3] A. Silberstein, A. Machanavajjhala, and R. Ramakrishnan. Feed following: the big data challenge in social applications. In *ACM SIGMOD Workshop on Databases and Social Networks (DBSocial)*, pages 1–6, 2011.