Data Integration in Social Networks – A Survey

CSci 5707 – Principles of Database Systems
Fall 2008

Project Guide: Prof. Jaideep Srivastava

by
Rasik Phalak [3908703]
Srinivasan Krishnan [3942424]
# Table of Contents

1. Abstract
2. Introduction
3. Disparities in data formats
   3.1. Types of heterogeneities in data
   3.2. Data models based on data formats
4. Data integration strategies
   4.1. Local-as-view (LAV) model
   4.2. Global-as-view (GAV) model
   4.3. Reliability Hierarchy (Egg/Yolk) model
   4.4. Hypergraph data model (HDM)
   4.5. P2P data models
5. Concepts in social networking
   5.1. Web services
   5.2. Swoogle: A search and metadata engine for semantic web
   5.3. FOAF
   5.4. Google OpenSocial
   5.5. Data Integration using the concept of Coreference
   5.6. XML
   5.7. OpenID 2.0
6. Challenges in data integration of social networks
   6.1. Issues in social networks
   6.2. Data integration issues in social networking
7. Conclusion
1. Abstract
In the modern world of technology, online communities and social networks have become the main channel for communication, fun and getting new acquaintances. The large number of proprietary social networks in the industry has resulted in vast but diverse information about people. In order to put them to commercial use, we need a one-point query interface for the information available among different social networks. Our paper deals with the analysis of the different strategies available for data integration, the concept of Semantic Web, representation of social networks using semantic web languages, and finally the issues faced while integrating social network data. We have also presented a brief study of the latest social networking developer interface called OpenSocial (by Google) and a unified login system for users called OpenID. These two initiatives have opened the door for resolution of various data integration issues in social networks.

2. Introduction
Data integration is the process of providing a unified view of the data spread across different sources. The main objective of data integration is to facilitate ease of information retrieval when queries are posed on a large set of sources. The most common data integration approach is to provide a set of materialized views over the source. The process of data integration plays a vital role in the fields of data warehousing and data mining.

A social network is a social structure in which the nodes are usually people or organizations, and the interlinks between them might be one among visions, ideas, friendship, kinship, dislike, trade, and so on. Social networks have become a popular area over the past few years, and have spread its roots from websites to online role playing games.

The availability of huge amounts of data in a social network tempts the commercial organizations to perform data mining and determine the patterns which might increase their revenue. In order to perform data mining, we need to have a unified view of data from different social networks. This is exactly the place where data integration of social networks becomes indispensable.

Semantic Web, which is called by Tim Berners Lee as a component of Web 3.0, plays an important part in integrating web based data. It facilitates data integration by making semantics of the web pages, machine understandable.

Attempts like Google OpenSocial have been made recently to provide application and web developers, a way to fetch and integrate data from various social networking websites. But they are still at their primitive stages of implementation and have a long road ahead.

In our survey, we present the prevailing data integration strategies in general and an overview of the technologies available to facilitate data integration in social networks. We also discuss the current prevailing issues in social networks and in integrating their data together.

3. Disparities in data formats
3.1. Types of heterogeneities in data
Heterogeneity is the major problem in data integration. In general, heterogeneity can be classified into three broad categories:

**Structural/Schematic Heterogeneity**
This arises because of the disparities in the schema of different databases. Most application programs require high speed data access, which might require denormalized schema and hence resulting in structural or schematic heterogeneity among the different data sources.

**Syntactic Heterogeneity**
This is the most common problem in the industry today, and it may arise because of the difference in protocols, languages/interface (SQL, ODBC, CORBA), data models (relational, object-oriented), etc.

**Semantic Heterogeneity**
This has become a more serious problem lately, and a lot of research work is in progress to resolve this. This type of heterogeneity arises because of the difference in the interpretation of meaning of the underlying data. Although dictionary based approaches are being used to mitigate the ill effects of this
heterogeneity, there are no hard-and-fast rules to completely eliminate this. The issue of semantic heterogeneity in data integration has not yet been conquered by the human brain.

3.2. Data models based on data formats
The data models based on the format of data, in general, can be demarcated into three main categories:

Structured data models
The model in which the data is well organized using some structure, like tables, is an example of a structured data model. They have well defined schema to represent information. The main advantage of this model is the easy with which queries can be imposed on the system. Examples of structured data include ER model, UML, relational model, etc.

Semi-structured data models
The data model in which there is no demarcation between the data and its schema is called a semi-structured data model. The nature of the application determines the number of attributes of the schema that are actually used. The main advantage of this system is that it provides a flexible format of data exchange between different types of databases, and that the schema can be easily changed. But on the downside, the queries on this system cannot be imposed as easily as in a structured model. Examples of semi-structured model include TCP/IP packets, Emails, zipped files, binary executables and all kinds of unstructured data.

Unstructured data models
Any kind of data which does not have a proper structure or schema comes under this category. Examples of this will include images and graphics, digital documents, etc. Any unstructured data is also a semi-structured data, because though they are just a stream of bits without a structure as such, the file as a whole carries some structure, so that the corresponding viewer can read and modify them.

4. Data Integration Strategies
New strategies to integrate data evolve from time to time, owing to the active research work being done in the field. However, not all strategies are applicable to all situations. The choice of the data integration strategy is a crucial decision to be made based on the type of data being targeted. The following are a few popular data integration strategies:

4.1. Global-as-view (GAV) Model
In this model, the global database is be modeled as a set of views over the source database. The advantage of this model is that the query processing is simpler. But the disadvantage is that the addition of a new source needs considerable effort. This makes GAV a good choice when the sources are less probable to change.
4.2. Local-as-view (LAV) Model
In this model, the source database is be modeled as a set of views over an underlying global schema. The advantage of this model is that new sources can be added easily when compared to GAV. However, the query rewriting process is complex because the system has to choose from a set of choices to determine the best possible rewrite.

![Diagram of Local-as-view (LAV) Model]

4.3. Reliability Hierarchy Model (Egg/Yolk model)
This model suggests the formation of mappings between the typical members of two classes. For example, let us assume that two companies form a merger and they have an employee database each. Let the different kind of employees in each firm be engineer, manager, and finance officer. There might be many instances of each of these kinds of employees for each company. According to this model, if we are able to determine the mapping between a subset of each category of employee (called as yolk), those mappings can be generalized across the whole set of employees (called as egg). The extent of overlap between the two yolks suggests the reliability of the mapping.

4.4. Hypergraph Data Model (HDM)
This is a primitive data model from which all higher level data models (either structured or semi-structured) can be derived. Data integration is achieved by means of the Hypergraph Query Language (HQL) and HQL views (mediators). All objects are characterized using nodes and hyper edges. The advantage of the model is its conceptual simplicity.

4.5. Peer-to-Peer Data Model (P2P)
A common approach to model P2P systems is to use a global LAV approach (GLAV). In this model, each peer has its own version of the global schema of P2P network. So the queries for any peer can be answered by any other peer in the network with a decent accuracy.

5. Concepts in Social Networking

5.1. Web services
Web services available on the internet which provide web API’s and are executed on a remote system hosting the web service. Using a web service, we can programmatically extract and integrate data from heterogeneous information systems. Currently, the following standards are being used for system integration. Technical integration challenges were solved by Web services by standardizing the infrastructure for data exchange.

Data communication between systems takes place in standard XML format.
SOAP – Simple Object Access Protocol is used to send and receive XML documents.
WSDL – Web Services Description Language is used for developing aggregation interfaces for web services.
UDDI – Universal Description, Discovery, and Integration is used to publish a registry of all system interfaces.
Different models have been proposed for aggregation using web services:

- **Content Aggregation** – Gathers content pertaining to a specific topic from varied sources and provides value added analytics based on relationships across multiple data sources.
- **Comparison Aggregation** – Based on the user specified criteria, compares results from various domains and provides the optimal results.
- **Relationship Aggregation** – Provides a common point of reference between a user and several business services/information sources with which the user has a business relationship.
- **Process Aggregation** – Business processes which require coordination across a variety of services/ information sources and managed and a common point of contact is provided.

Challenges:

- Transformation amongst different meanings attached to same standards, names in web services paradigm is often the most difficult integration challenge to overcome. E.g. Bandwidth units used by different web services might be different.
- Modularization of Business processes like EIS applications is functionally difficult to implement.
- Secure access of web services and authenticity of a web service to a user cannot be guaranteed.
- The quality, accuracy, consistency and correctness of information provided by a Web Service cannot be guaranteed. Licensing and payment issues for a web services also needs significant user attention.
- Extracting meaning from text is the most challenging task for computer programs. Semantic web is an avenue for encoding and publishing information in ways that makes it easier for computers to understand and interpret the information.

5.2. Swoogle: A search and metadata engine for semantic web

A semantic web document (SWD) is a document in a semantic web language that is online and accessible to web users and software agents. SWD is classified further as Semantic Web Ontologies (SWO’s) and Semantic Web Databases (SWDB’s). SWO’s are fundamentally used for definitions of new terms and extending the definitions of existing terms used in a SWD. SWDB’s do not introduce new terms or extend definitions about those terms, but introduce individuals and make assertions about individuals defined in other SWD’s.

Swoogle uses a web crawler to search URL’s using the Google Web Service that discovers all the semantic web documents, a metadata generator and a database that stores metadata about the discovered semantic web documents.

Upon discovery of relevant information from the web, Swoogle uses methods like imports(A,B), uses-term(A,B), extends(A,B), asserts(A,B) which links the two SWD’s A and B and uses a ranking function to establish the relevance between the two documents. With the advent of web 3.0, most of the information on the web would be represented using SWD’s. Using SWD’s it is possible to integrate the human readable and machine understandable information together fetch more relevant results for the user query. The classification of documents as SWO’s and SWDB’s is an important issue which needs more attention with respect to appropriate data integration. After appropriate classification, different SWD’s can be merged together to display information gathered from various sources in response to the user query.

5.3. FOAF

Various social networks allow their members to publish their profile information including their social links, using Resource Description format (RDF). The RDF vocabulary defined by FOAF – Friend of a Friend Ontology.

FOAF vocabulary includes classes and properties useful to which are used to describe people online. FOAF defined 12 classes and 51 properties which are used to construct the basic social networks. Using FOAF for information gathering in social networks consists of Identification of FOAF documents, extraction of person information and fusion of person information based on the semantics of FOAF vocabulary. The variety and richness in the information which can be represented using FOAF vocabulary allows to identify social ties and identify friendship types. FOAF acts as a bridge for gathering some extra information about individuals which can link them like research interests, photos.
shot together. Some FOAF properties can be declared as ‘inverse functional’, which can help us to identify if 2 individual FOAF nodes represent the same person. Fusion of information about a person spread across the social network can be done as:

- foaf:person can be identified as describing the same person as
  - Two anonymous individuals sharing the same URIref in RDF graph can be merged together to represent the same person.
  - Using Web ontology language inverse functional property, assertions can be made if the two individuals identified are the same.

However using these techniques does not guarantee complete accuracy in the results derived, and hence appropriate care must be taken while merging information from multiple FOAF documents.

Example: Suppose the email ID of a person is mistakenly written as phal003@umn.edu instead of phala003@umn.edu, and both the email ID’s actually exist, then this would lead to merger of individuals who are not linked together in any way.

5.4. Google OpenSocial
Google OpenSocial is an platform that allows to merge information about distinct individuals by enabling social networks to interlink and self organize into a social ecosystem guided by the policies of individuals and organizations. Using Google open social for interlinking information about individuals from various social networks may lead to unsolicited interaction within the alliance formed between the communities offering the social network. If appropriate measures are taken for avoiding this, then open social aims to transform the internet from a provider centric (multiple consumers related to a provider) to a customer centric (multiple providers are related to a consumer) paradigm. Identity providers can be used to validate the authenticity of individuals/organizations on the social network.

OpenSocial uses a two-fold approach for establishing connections amongst entities on the social network.
- Initially the discovery stage determines the entities that are suitable to establish a connection with for future interactions.
- After the initial discovery process, the interaction activity process is initiated which leads to the interaction coordinator to negotiate with the other entity on the interaction activity policy which enables linking and interaction between entities on the social network.

5.5. Data Integration using the concept of Coreference
Coreference is used to describe the situation where different terms are used to describe the same referent. When records and files from different databases are merged together, elimination of duplicates during information integration must be considered. Similar concept applies to merging of information about individuals on a social network.

If information about an entity is derived from multiple resources/URI then Web ontology language provides the sameAs attribute which can be used to establish a link between these similar information sources and bundle them together into a group. The semantics of the sameAs attribute specifies that URI’s linked with a particular predicate have the same identity. This approach might sometimes lead to ambiguities as two URI’s might actually refer to 2 different entities; however they are merged together based on the context in which they are specified.

The Consistent Reference Service (CRS) has been specified to manage coreference between millions of URI’s accumulating over the internet. It is implemented using both an RDF knowledge base and a relational database using RDF export.

5.6. XML
Currently most of the data over the Web is represented in XML format as it provides greater flexibility in the kinds of data that can be handled. Various models for integration of XML data have been proposed, however the following challenges still encircle the issue of integration. Data Models for XML data integration should also provide support for other common formats which are frequently
used by the users like hierarchical, relational, etc. Also data integration form multiple data models should also be supported. Integration of XML data from a document with data provided by a model which is used for XML data integration should also be supported.

5.7. OpenID 2.0

It provides a common specification for logging into multiple sites. Whenever the user logs in for the first time, the query is redirected from the website the user is trying to authenticate to a claimed identifier. The claimed identifier returns the URI of the user’s OpenID authentication service endpoint. The website, then communicates with the Identity Provider to create a Shared Secret identity for that particular user. This secret identity is then redirected from the website to the user via the new URL which redirects the user to the identity provider. The user logs in at the Identity Provider and completes the trust authentication process. The Identity Provider then redirects the user to the Website with the proof that the user is authentic and the URL is owned by the user. The Identity Provider also provides any profile data which the user has agreed to be made public. The user is now logged in to the website after completing the authentication process.

6. Challenges in data integration of social networks

6.1. Issues in social networking

- The primary motive of any business entity is to gain monetary gains and social networks are not an exception to it. From the company’s perspective who is developing the website, monetary gains are not that significant as statistics show that only 4 of 10000 users click on adds which is the primary source of income.
- One additional factor to this is that social networking sites have banned at work places which reduces their popularity.
- Since sending information on a social network to peers is easy and no restriction are exist on the same, social networking spam is an important issue into consideration.
- Forged identity on the social network is extremely difficult to track and hence its prevention is also difficult.
- Users can no longer trust the sophisticated applications which are published over the network.
- Once information is published on any social network, due to the presence of web crawlers, the information can immediately get propagated across the web. Hence deletion of information by the user in the future does not result in complete elimination of information from the web.

6.2. Data integration issues in social networking

- Social networks do not pose any restriction in terms of the amount of data which a user can publish over the web. Since massive amount of data is available on the network, integration amongst it is an issue.
- Privacy and security concerns for the data published over the web still remain and there are no means for monitoring any unauthorized access to data in social networks. Trust amongst the individuals is an important consideration in this respect, however if a trusted individual turns malicious, there are no means for detecting such cases.
Misrepresentation of information in social network leads to incorrect FOAF mappings. Slight inaccuracy in the information recorded in FOAF documents leads to unnecessary chaos and linking amongst incorrect entities on the social network.

In spite of data integration at the core, the issue of multiple login for different social networks still haunts the user.

Dynamic application updates may result in unwarranted access to private information. No measures have been undertaken to prevent dynamic application updates in any social forum.

7. Conclusion

Data integration in social networks is a new kid on the block and is generating a lot of interest amongst the researchers. This might be due to the potentially high commercial value attached to the development of Social networks. The next generation web pages, also called as semantic web documents (SWDs), make the task of integration easier by merging human-friendly representation of information with machine-friendly representation of data. Based on the survey performed, it can be concluded that FOAF representation with RDF would be the most popular choice for the representation of data in social networks in future. Google's initiative of OpenSocial, which provides an API to integrate different social networks, combined with OpenID 2.0, which provides a unified login for different websites, has a good probability of becoming a great success in the field of social networking. We have also presented the various privacy and integrity issues in the process of data integration in social networking.
References

1. Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection
   Boanerges Aleman-Meza, Meenakshi Nagarajan, Cartic Ramakrishnan, Li Ding, Pranam Kolari, Amit P. Sheth, I. Budak Arpinar, Anupam Joshi, Tim Finin

2. Social networking on the semantic web
   Tim Finin, Li Ding, Lina Zhou and Anupam Joshi

   Li Ding Tim Finin Anupam Joshi Rong Pan R. Scott Cost Yun Peng, Pavan Reddivari Vishal Doshi
   Joel Sachs

4. Spinning Multiple Social Networks for SemanticWeb
   Yutaka Matsuo and Masahiro Hamasaki and Yoshiyuki Nakamura, Takuichi Nishimura and Koiti Hasida

5. POLYPHONET: An Advanced Social Network Extraction System from the Web
   Yutaka Matsuo, Junichiro Mori, Masahiro Hamasaki

6. Data Integration Using Web Services
   Mark Hansen, Stuart Madnick, Michael Siegel

7. Web-scale Data Integration: You can only afford to Pay As You Go
   Jayant Madhavan, Shawn R. Jeffery, Shirley Cohen, Xin (Luna) Dong, David Ko, Cong Yu, Alon Halevy

8. URI Identity Management for Semantic Web Data Integration and Linkage
   Afraz Jaffri, Hugh Glaser and Ian Millard

9. OpenID 2.0: A Platform for User-Centric Identity Management
   David Recordon and Drummond Reed

10. OpenSocial: From Social Networks to Social Ecosystem
    Juliana Mitchell-Wong, Ryszard Kowalczyk, Albena Roshelova, Bruce Joy and Henry Tsai

11. Integration of Semantic Data Using a Novel Web Based Information Query System
    Okkyung Choi, Sangyong Han and Ajith Abraham

12. The Nimble XML Data Integration System
    Denise Draper, Alon Y. Halevy, Daniel S. Weld

13. Semantic Data Integration in P2P Systems
    - Diego Calvanese, Elio Damaggio, Giuseppe De Giacomo, Maurizio Lenzerini and Riccardo Rosati

14. The Egg/Yolk reliability hierarchy: Semantic Data Integration Using Sorts with Prototypes
    - Fritz Lehman and Anthony G. Cohn

15. Semantic integration and querying of heterogenous data sources using a hypergraph data model
    - Dimitri Theodoratos

16. Spinning Multiple Social Networks for SemanticWeb
    - Yutaka Matsuo and Masahiro Hamasaki and Yoshiyuki Nakamura, Takuichi Nishimura and Koiti Hasida