Social Sifter: An Agent-Based Recommender System to Mine the Social Web

M. Omar Nachawati, Rasheed Rabbi, Genong (Eugene) Yu, Larry Kerschberg and Alexander Brodsky

Dept. of Computer Science
George Mason University
Fairfax, VA, USA
{mnachawa, rrabbi, gyu, kersch, brodsky} at gmu.edu

Abstract**—** With the recent growth of the Social Web, an emerging challenge is how we can integrate information from the heterogeneity of current Social Web sites to improve semantic access to the information and knowledge across the entire World Wide Web, the Web. Interoperability across the Social Web sites make the simplest of inferences based on data from different sites challenging. Even if such data were interoperable across multiple Social Web sites, the ability of meaningful inferences of a collective intelligence [1] system depends on both its ability to marshal such semantic data, as well as its ability to accurately understand and precisely respond to queries from its users. This paper presents the architecture for Social Sifter, an agent-based, collective intelligence system for assimilating information and knowledge across the Social Web. A health recommender system prototype was developed using the Social Sifter architecture, which recommends treatments, prevention advice, therapies for ailments, and doctors and hospitals based on shared experiences available on the Social Web.

Keywords: social semantic search; collective knowledge systems; recommender systems, OWL; RDF; SPARQL

I. INTRODUCTION

Since its inception, the World Wide Web has always overwhelmed users with its vast quantity of information. The advent of Social Webs, coined Web 2.0, has placed an additional burden on Web search engines. While the established algorithms that Web search engines employ are effective in surfacing the most popular results through hyperlink analysis, as demonstrated by the Hubs and Authorities algorithm [2] and the PageRank algorithm [3], those results are not necessarily relevant despite popularity and these algorithms have fallen short of solving the problem of information overload [1, 2, 3] on the World Wide Web.

The research into natural language understanding [4] attempts to close that gap. However the quality of machine generated semantics still pales in comparison to that of humans. This became a core challenge for the Semantic Web or Web 3.0, where information is made available in structured, machine-friendly formats allowing machines not only to sort and filter such data, but also to combine data from multiple Web sites in a meaningful way and allow inferences to be made upon that data. While semantic query languages, such as SPARQL, can provide a database-like interface to the World Wide Web, it is only as good as the quantity and quality of information that is made available in structured, machine readable formats, such as RDF and OWL .

Conventionally, finding answers to questions and learning from the knowledge mine existed on the Social Web has primarily been a manual process. It requires a lot of intelligence in sifting through the mountains of Social Web pages using only a keyword-based Web search engine, which is akin to a primitive pitch-fork in Semantic Web terms. More recently, however, Social Web sites have begun to embrace Semantic Web technologies such as RDF and OWL, and have been offering much more machine-friendly data, such as geo-tagged images on Flickr, Friend Of A Friend (FOAF) exports in FaceBook and hCalendar [7] tagged events on Blogger. Such developments have sparked the evolution of the Social Web into a collective knowledge system [1], where the contributions of the user community are aggregated and marshaled with knowledge from other heterogeneous sources (e.g., web pages, news and encyclopedia articles, and academic journals) in a synergy dubbed the *Social Semantic Web*.

While the Semantic Web focuses on data to enable interoperability among heterogeneous semi-structured web pages, the focus of the Social Semantic Web vision is to create a system of collective intelligence by improving the way people share and explore their own and others knowledge and experience [1]. Work on the Social Sifter promotes that grand vision and expands on the research done on the patented Knowledge Sifter architecture [7, 8, 9], as well as the Personal Health Explorer [11], undertaken at George Mason University. As a proof of concept, we have designed a social health knowledge and recommender system based on the Social Sifter platform that utilizes the Social Semantic Web to provide precise search results and recommendations.

The rest of this paper is organized as follows: section II discusses related work, section III describes the Social Sifter architecture and a brief description of the prototype system. Section IV highlights the experimental results, and Section V identifies the possible future work on the Social Sifter platform.

II. RELATED WORK

A. Knowledge Sifter and Personal Health Explorer

Semantic systems belong to a class of systems that make use of ontologies, context awareness and other semantic methods to make informed recommendations. Such research in semantic search at George Mason University began with WebSifter [8, 9, 10], an agent-based multi-criteria ranking system to select semantically meaningful Web pages from multiple search engines such as Google, Yahoo, etc. The work further led to a patent [8]. Knowledge Sifter (KS) [8] is motivated by WebSifter [7,8], but is augmented with the advanced use of semantic web ontologies, authoritative sources, and a service-oriented plug-and-play architecture. Knowledge Sifter is a scalable agent-based web services framework that is aimed to support i) ontology guided semantic searches, ii) refine searches based on relevant feedback, and iii) accessing heterogeneous data sources via agent-based knowledge services. Personal Health Explorer (PHE) is an enhancement of KS to perform semantic search in biomedical domain. PHE leverages additional features of a personal health graph to be identified, categorized, and reconstituted by providing links to the user to rate individual results and return to previous queries and update information through a semantically supported path.

KS and PHE are able to obtain more relevant search results than classic search engines; while the result is very general, it leaves room to make it more personalized. Both KS and PHE make multifaceted efforts towards realizing the Semantic Web vision, primarily focusing on the formal ontological sources. PHE provides facilities to include a user's Personal Health Record (PHR), which entails additional permission and access control which may be constrained by HIPAA regulations. Interestingly, both of these systems did not use the data available on the Social Web, namely Wikipedia, YouTube, Flickr, Facebook, LinkedIn, etc. This is where Social Sifter makes its contribution.

B. *BLISS and Cobot*

Other attempts to utilize Web 2.0 technology to enhance the quality and relevance of health recommendation systems include bookmarking, crowd sourcing, crowd tagging and harvesting user recommendations. The Biological Literature Social Ranking System (BLISS) is one such prototype system that allows users to bookmark and promote their recommendation to communities of special interest, facilitate the annotation and ranking by the community, and present the results to allow other users to get the recommendations based on community ranking [6]. The bookmarking approach is useful in establishing the authoritativeness of information over the long term because it uses social voting or ranking [5].

The Cobot system uses social conversation and social tagging (preference) to enhance the health recommendations. Three techniques are noteworthy: (1) user-initiative dialogue in capturing user's intent, (2) social tagging in establishing the authoritativeness of social information, and (3) case-based semantic reasoning in utilizing social knowledge for recommendation [5].

C. *Semantic Analytics on Social Networks*

A multi-step engineering process is described in [9] to utilize social knowledge. These steps are common procedure to across the initiatives to transform the social web information to semantic knowledge.

Social Sifter adheres to the underlying framework of Knowledge Sifter [9], the knowledge manipulation mechanism of PHE [10], and engineering process for semantic association of [11] to leverage an integrated semantic search engine and recommender system.

III. The Social Sifter Architecture

Social Sifter, an enhancement of the existing Knowledge Sifter (KS), is a collection of cooperating agents that are exposed through web services and exhibits a Service-Oriented Architecture (SOA)-based framework.

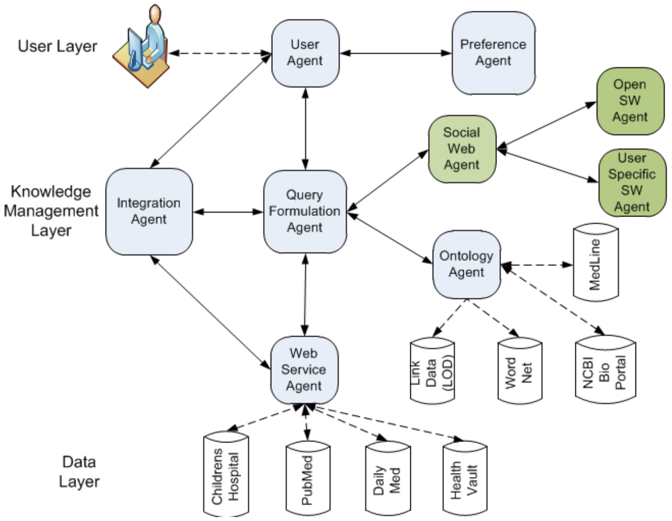

Figure 1. Social Sifter Architecture – Tiers and Components

Depending on the functionality, agents are allocated into three different architecture layers – i) the User Layer, ii) the Knowledge Management Layer, and iii) Data Layer. The User Layer consists of the User and Preferences agents, and manages all user interaction and data preferences. The Knowledge Management Layer handles the support for semantic search, access to data sources, and the ranking of search results using technologies like the Ontology, Social Web Crawling, Ranking, Query Formulation, and Web Services agents. The Data Layer consists of the data repositories that provide authoritative information and documents. The hierarchy of the architecture layers is already defined in KS; three additional agents were added, with an alteration of the underlying algorithm to perform the execution flow into the Social Sifter.

Social Web agent basically collaborates with following two agents to manipulate social web information.

Open SW agent performs open search within the blogs, related support groups etc.

User Specific SW agent identifies user social identities across the web and conducts Collaborative Filtering by processing social tags, user participation and responses available on the social webs.

IV. Health Recommender System

As a proof-of-concept, we are building a health recommender system using our Social Sifter architecture that provides health recommendations for any type of sickness, disease or disorder. The present system does not do any natural language processing on user queries, and therefore is limited as to what it can accept as a valid query. Currently, the system accepts a comma delimited list of words that relate to a specific ailment and returns a list of relevant descriptions of the ailment, therapy options, doctors, and treatment centers as collected from the Social Semantic Web from our knowledge Management Layer. We intend for future versions of the health recommender system to allow for unrestricted language queries by performing natural language processing to transform the unstructured query input into a more structured format, acceptable by the Social Sifter architecture.

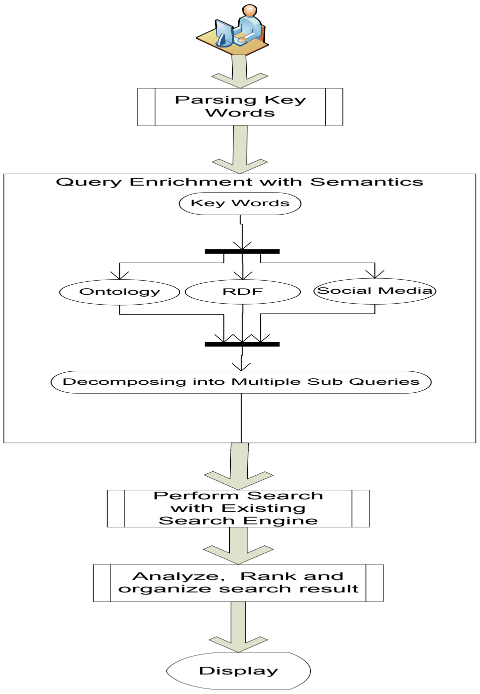
Parsing Key Words
Query Enrichment with Semantics
Key Words
Ontology
RDF
Social Media
Decomposing into Multiple Sub Queries
Perform Search with Existing Search Engine
Analyze, Rank and organize search result
Display

Figure 2. Social Sifter work flow diagram

A. Scenario for Pancreatic Cancer

Consider the case when a user is exploring recommendations for *pancreatic cancer*. According to the NIH, treatment options include surgery and biliary stents. The NIH also lists links to support groups, among which CancerCare.org features a social question-answer forum that is categorized by topic. Our inference agent for health recommendations takes advantage of this domain knowledge in attempting to provide better quality recommendations than what would be available from a general Web search engine. Let us walk through the steps of the health recommender system for this particular query.

Query Submission: User logs into the health recommender system website and enters the following query terms "pancreatic cancer."

Query String Preparation:

i) The ***User Agent*** parses the query string to identify key words.

ii) The ***Preference Agent*** collects context information, including the user's IP address, a query session identifier, and the best geographic location estimate available for that user. It tries to create a User Profile by indexing friendship and affiliation information to generate the user's Social Graph.

iii) User Agent passes the SPARQL query and the collected User Profile information to the Query Formulation Agent.

Query Refinement: The ***Query Formulation Agent*** then attempts to enrich the original SPARQL query by:

i) ***Semantic Query Decomposition***: It will generate multiple sub-queries that generalize and specialize the term *pancreatic cancer* based on the health-domain ontology from The National Center for Biomedical Ontologies (NCBO), a BioPortal and MedLine (Medical Literature Analysis and Retrieval System Online), which is a bibliographic database of life sciences and biomedical information.

ii) ***Marshalling***: selected data will be marshaled with the amassed folksonomy from the Social Web Agent. The inference engine will also generate queries based on the results of any cluster analysis from data crawled from the Social Web, which may pick up, for instance, other ailments that people have discussed together with *pancreatic cancer*.

iii) ***Ranking***: The end result of this meta-search is a weighted tree of sub-queries, where weights are assigned based, among other features, on the static nature of the sub-query generated (heuristically) as well as the importance of the source (back-reference analysis).

Post Query Processing: Once all sub-queries have been defined, the Web Service Agent passes them to the Data Layer, which accordingly runs the queries and itself ranks each result, based on many factors, including relevance (ontological), importance (back-reference based) and belief (Bayesian-based inference from Social Semantic Web).

Result Scrutinizing: The results are then returned to the Integration Agent, which combines different classes (based on the results from the classifier) of results based on a total ordering derived from the aggregated ontology, and back-reference analysis. The agent also performs a clustering analysis on the result set to further group the results and perform statistical calculations on the groups of results before passing them to the User Layer.

Result displaying: The User Layer then displays the grouped and ranked results according to the preferences selected by the user.

B. Query life cycle for Pancreatic Cancer in Social sifter

The life cycle of a query in Social Sifter, e.g., searching for "pancreatic cancer", is as follows: (1) a user allows access to his profile, (2) Sifter culls information from his social networks, (3) Sifter initiates targeted information harvesting, (4) Sifter conducts semantic inference and reasoning, and (5) Sifter presents socially- and semantically-renked results are to the user.

C. Social Sifter Prototype

The Social Sifter prototype has been implemented to use information retrievable from Facebook using Graph API in gathering the information about the users. In Facebook, each user can have feeds, likes, activities, interests, music, books, videos, events, groups, checkins, games, and his personal information, like hometown and related locations. These provide a very rich base for understanding the intension of a user when he is searching on the Web.

Social Sifter combined both semantic reasoning and social ranking to better understand user's intention and present the results to users, based on initial search keywords or phrases provided. The algorithm for the currently implemented search is described as follows.

(1) Login: User logs into his Facebook using OAuth authentication. The program gets the authorized token and uses it to access user's information with user's concurrence.

(2) Information Retrieval: The system retrieves the information about the user (Feeds, Likes, Activities, Interests, Music, Books, Photos, Videos etc.) and uses them in supporting the targeted harvesting of information and formulating the social ranking of results in categories.

(3) Social ranking – A simple algorithm is used to calculate the social weights of the harvested information in each category. The algorithm is basically counting the occurrences of keywords or phrases in each category.

(4) Social context – The user's background information is used in refining the search results or filtering the results. One specific example is the location information. The home location of the person is generally used to limit the places to be searched and returned.

(5) Semantic result presentation – The results are presented to users in groups: people, groups, events, places, events, pages, or posts. The current implementation is limited to use the categories or semantics of Facebook. The actions in Facebook link objects and people. They are the bases for our search engine in weighing the harvesting strategies. They are also important in ranking the results and the categories when presenting the search results to users. The current implementation used the same social ranking strategy described in (3).

D. Proactive Social Search

The existing Facebook semantics do not capture the semantic of health queries. For health problems, users may be interested in finding out the cure of certain diseases, which is not captured by the current set of actions available in Facebook. Customized actions can be implemented using the Facebook Open Graph, but it is beyond the scope of this paper.

V. Expermental Findings

The Social Sifter prototype has been implemented. The Facebook Graph API was used as the basis for harvesting social network information about the user. Social information was used in two aspects – understanding the user's intention (context) and ranking results (social semantic ranking). The two aspects showed improved search results. For example, the searching case using phrase – "pancreatic cancer" can be compared using three different engines – Google, Facebook, and Social Sifter. Social Sifter provided integrated results and used social ranking to rearrange the categories depending on users profile information. Location is determined based on user provided current living locations. More testing is being carried out to determine metrics to assess the quality of social semantic search recommendations.

VI. Conclusions

Social semantic search is an integration of social networks and semantic search. Semantic search provides rich means in enhancing search, especially the user's intent and semantic reasoning. Social search involves people and links to their social graphs. In this paper, a prototype social semantic search engine, Social Sifter, has been presented. The lessons learned from the implementation showed two areas for improving search accuracy: social contextual information (user intent understanding) and social semantic ranking (results relevance).

The current implemented prototype system is limited in the use of the semantic reasoning. The crawling of data should be expanded to other social media and social networks. Integration of these results into a standard semantic data store is necessary to realize the power of semantic reasoning. Further study directions are: (1) to integrate mature ontologies, (2) to define customized actions to demonstrate the approach in health domain, and (3) to use the reasoning power of semantics.

References

[1] T. Gruber, "Collective knowledge systems: Where the Social Web meets the Semantic Web," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 1, pp. 4–13, Feb. 2008.

[2] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*. 1999.

[4] A. Ntoulas, G. Chao, and J. Cho, "The infocious web search engine: improving web searching through linguistic analysis," in *Special interest tracks and posters of the 14th international conference on World Wide Web*, New York, NY, USA, 2005, pp. 840–849.

[5] J. M. Gomez, G. Alor-Hernandez, R. Posada-Gomez, M. A. Abud-Figueroa, and A. Garcia-Crespo, "SITIO: A Social Semantic Recommendation Platform," in 17th International Conference on Electronics, Communications and Computers, 2007. CONIELECOMP '07, 2007, p. 29–29

[6] "hCalendar 1.0 · Microformats Wiki." [Online]. Available: http://microformats.org/wiki/hcalendar. [Accessed: 15-Apr-2012].

[7] L. Kerschberg, W. Kim, and A. Scime, "WebSifter II: A Personalizable Meta-Search Agent Based on Weighted Semantic Taxonomy Tree," in *International Conference on Internet Computing*, Las Vegas, NV, 2001

[8] L. Kerschberg, W. Kim, and A. Scime, "Personalizable semantic taxonomy-based search agent," U.S. Patent 7117207Oct-2006

[9] L. Kerschberg, H. Jeong, Y. Song, and W. Kim, "A Case-Based Framework for Collaborative Semantic Search in Knowledge Sifter," *Case-Based Reasoning Research and Development*, vol. 4626/2007, pp. 16–30, 2007.

[10] T. G. Morrell and L. Kerschberg, "Personal Health Explorer: A Semantic Health Recommendation System," workshop on Data-Driven Decision Support and Guidance System (DGSS), 28th IEEE International Conference on Data Engineering, Arlington, VA April 1, 2012.

[11] Boanerges Aleman-Meza, Meenakshi Nagarajan, Cartic Ramakrishnan, Li Ding, Pranam Kolari, Amit P. Sheth, I. Budak Arpinar, Anupam Joshi, Tim Finin. Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection. *WWW 2006,* May 23–26, 2006, Edinburgh, Scotland. ACM 1-59593-323-9/06/0005.