

Socio-Semantic Health Information Access

Saurav Sahay

College of Computing
Georgia Institute of Technology
Atlanta, Georgia

Ashwin Ram

College of Computing
Georgia Institute of Technology
Atlanta, Georgia

Abstract

We describe Cobot, a mixed initiative socio-semantic conversational search and recommendation system for finding health information. With Cobot, users can start a real time conversation about their health concerns. Cobot then connects relevant users together in the conversation also providing contextual recommendations relevant to the conversation. Conventional search engines and content portals provide a solitary search experience inundating the health information seeker with a hoard of information often confusing and frustrating them. Cobot brings relevant healthcare information directly or through other users without any search through natural language conversation.

Introduction

Search Technologies: An Evolution

Search engine technologies are a practical application of information retrieval (IR) to large-scale document collections. With significant advances in computers and communications technologies, people today have interactive access to enormous amounts of user-generated content on the Web. This has spurred rapid growth in search engine technology, where search engines are trying to discover different kinds of entities such as users, discussions, answers to questions or other precise information nuggets found on the Web with emphasis on real time information access.

Semantic approaches to IR use knowledge-based techniques of retrieval that broadly rely on the syntactic, lexical, sentential and discourse-based levels of knowledge understanding. Semantic approaches include different levels of analysis, such as morphological, syntactic, and semantic analysis, to model, extract and reason from information sources more effectively. The development of a sophisticated semantic system requires complex knowledge bases of semantic information as well as retrieval heuristics. There are a few natural language search engines such as Hakia¹ and Powerset² (now part of Bing) that aim to understand the structure and meaning of queries written in natural language text, generally as a question or narrative.

Copyright © 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹www.hakia.com

²www.powerset.com

Agent-based approaches (Chen and Sycara 1998) involve the development of sophisticated artificial intelligence systems that can act autonomously or semi-autonomously on behalf of a particular user, discover and process information, e.g. (Allen et al. 2007). Intelligent Web based software agents search for relevant information using characteristics of a particular domain to organize and interpret discovered information. Personalized Web agents are another type of Web agents that utilize the personal preferences of users to organize search results, or to discover information that could be of value for a particular user. User preferences could be learned from previous user choices, or socially from other individuals who are considered to have similar preferences to the user. Cobot is a socio-semantic intelligent information agent for health communities that analyzes conversations and social preferences to provide socially filtered, semantically analyzed conversational recommendations.

Information seeking is mostly a solitary activity on the Web today. The traditional view of Web navigation and browsing assumes that a single user is searching for information. This view contrasts with previous research by library scientists who studied users information seeking habits. Recent research has demonstrated that additional individuals may be valuable information resources during information search by a single user. Studies have shown that there is often direct user cooperation during Web-based information search. Some studies report that significant segments of the user population are engaged in explicit collaboration on joint search tasks on the Web. Active collaboration by multiple parties also occur in certain cases; at other times, and perhaps for a majority of searches, users often interact with others remotely, asynchronously, and even involuntarily and implicitly. Socially enabled online information search (social search) is a new phenomenon facilitated by recent Web technologies. (Horowitz and Kamvar 2010) Collaborative social search involves different ways for active involvement in search related activities such as co-located search, remote collaboration on search tasks, use of social network for search, use of expertise networks, involving social data mining or collective intelligence to improve the search process and even social interactions to facilitate information seeking and sense making. Social psychologists have experimentally validated that the act of social discussions has facilitated cognitive performance (Ybarra et al. 2008). People in

social groups can provide solutions (answers to questions), pointers to databases or to other people (Cross, Rice, and Parker 2001)(Fox et al. 1993) (meta-knowledge), validation and legitimization of ideas (Evans and Chi 2008), and can serve as memory aids (Karasavvidis 2002) and help with problem reformulation. Guided participation is a process in which people co-construct knowledge in concert with peers in their community. One of the goals of the Cobot system is to actively engage users for collaborative problem solving during the conversational search activity.

Cobot

In this paper, we give an overview of Cobot, a socio-semantic conversational search system for recommending contextually relevant users and documents in conversations (Figure 1).

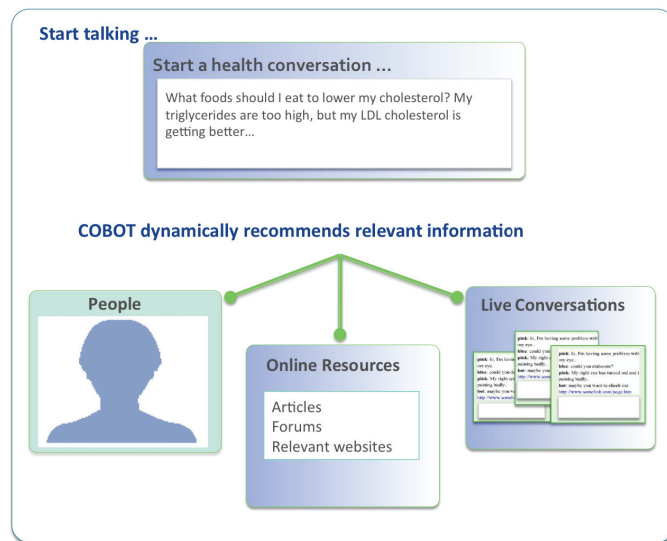


Figure 1: Cobot Functionality

Overview

Main Components

The main components of Cobot system can be classified into the following:

1. *Language Understanding*
2. *User Modeling*
3. *Case based reasoning*

We briefly describe each component of Cobot system in the following sections.

Language Understanding

Intent Detection Conversational interactions are classified into one of the following categories in Cobot to strategize for query reformulation stage and to help make the decision if the agent should insert some type of recommendation into the conversation:

- **ASK QUESTION:** Asking a question, e.g. somebody posts a problem. This is usually, but not always, the first post of a thread.
- **DITTO:** Repeating a question, e.g. "Yes, I also have the same (or a very similar) problem".
- **ASK CLARIFICATION:** Asking for more details about the problem, e.g. "Can you please provide more details?"
- **FURTHER DETAILS:** The person who is facing a problem provides more detailed information about it, possibly after somebody asks for more details.
- **SUGGEST SOLUTION:** Suggesting a solution
- **EXPRESSIVE** (Thanks for suggestion/solution, complaints about suggestion/solution, reject/accept solution)
- **OTHER** (Not fitting the above categories)

We have trained our Intent Detection classifier by annotating and training WebMD conversations threads. The accuracy of our classifier is close to 70% (Sahay and Ram 2010).

Query Generation Cobot analyzes conversations to extract concepts, relationships between concepts and focus of conversations to generate meaningful queries for external search engines for bringing in relevant candidate results. We use OpenNLP chunker trained on medical corpus (E. Buyko 2006) to extract phrases and map them into concepts using UMLS ontology (Aronson 2001). Main concepts expanded with their synonyms in conversations help us in retrieving recall oriented documents. We extract Subject-Verb-Object (SVO) triples (Sahay et al. 2008) from sentences as queries to retrieve documents that closely match the context in conversations. We are also experimenting with generation of queries based on the predicate argument structure in sentences using ASSERT semantic role labeling system (Pradhan et al. 2004).

Semantic Tagging Socially enabled systems have the property of self-governance and evolution by its community. While the major challenge remains getting a critical mass, these require lesser coordination. The problem with social tagging is that the noise-signal ratio becomes high due to informal nature of the language in conversations. Cobot system normalizes these conversations to extract meaningful conceptual representations using the extensive UMLS ontology and approximate string matching to guide social tagging of conversations. Cobot's internal knowledge representation system uses the concepts from UMLS and Wordnet as its language of representation. Cobot also uses broad category topic classifiers to categorize text into medical categories - this is specially useful in the ranking stage where similar class items are given higher weights in ranking.

Ontology based reasoning Ontology based reasoning helps quantify how similar different concepts are by determining their conceptual distances in a hierarchy. UMLS links concepts to semantic types and they provide complete hierarchical paths of known concepts to the root of the UMLS trees. Cobot computes paths and distances between different concepts that occur in UMLS ontology for mapping

inter-document similarity. We use UMLS-Similarity package that implements some semantic distance metrics to compute the scores between two terms. (McInnes, Pedersen, and Pakhomov 2009)

User Modeling

Language and interaction (percepts) creates usable memories, useful for making decisions about what actions to take and what information to retain. Cobot leverages these interactions to maintain users' episodic and long term semantic models, agent's per conversation working memory of concepts, syntactic and semantic information nuggets, participating users and messages. The agent analyzes these memory structures to bring in external recommendations into the system by matching with the contextual information need. The social feedback on the recommendations are registered in the indices for the algorithms to generate their user specific and conversation specific contextual relevance.

The purpose of Episodic Memory is to capture the user's short-term interactions and interests. Based on user's frequency of interactions and diversity in topics, this memory empirically varies in the range of a few days for different users. The Semantic Memory captures the user's long-term profile. These are the topics that interest the user in general and for a prolonged time. These interests change less frequently and represent general criteria of recommendation to the user. Many times, users might be interested in some temporary information need. Such information need not be incorporated in the long term user memory. The episodic memory captures such short-term interests. The episodic memory forms a sort of staging area and the concepts from this memory are selectively and periodically moved to the semantic memory in a crossover process.

The nodes of the semantic memory are concepts extracted from user's interactions. The concepts are connected with associations which develop when concepts co-occur frequently. Over a period of time when the user participates in more interactions, new concepts are added to the semantic memory. Our system currently tries to find a recently active user first who participated in similar conversations. Different conversational facets are matched with episodic memories and a spreading activation search on the semantic net is performed for recommending the top 3 users for the conversation. The activation is spread to the neighboring nodes proportional to the weight of each connecting association in the semantic net. There are several parameters in the system that can be learnt based on activity of users. Parameters for episodic memory window size, semantic memory learning and unlearning rates, concept co-occurrences and feedback strengths for associations are initially set heuristically and can be fine-tuned to suit individual users.

Case based Reasoning

For web search and conversation recommendations, we reformulate queries from the conversation snippets based on occurrence of concepts, relationships, categories and types of intent in the conversations. For a given target query Q_t , similar past recommendations and conversations are ranked so that the results which are most likely related to the learned

preferences of the community are promoted (Smyth et al. 2009) (Pazzani and Billsus 2007) (McCarthy et al. 2006). This kind of personalization is based on the reuse of previous search episodes: the promotions for Q_t are those results that have been previously selected by community members for queries that are similar to Q_t .

Cases are represented as tuples made up of the query component (a set of query terms, Q_i used during some previous search session) along with web recommendations and past conversations with their community hit counts. Our formulation is based on similar work reported in Paper (Smyth et al. 2009). Each case is a summary of the community's search experience relative to a given query. Each new target problem (corresponding to a new query Q_t) is used to identify a set of similar cases in the case base by using a term-overlap similarity metric to select the n most similar search cases for Q_t . Relevance of a result with respect to the current target query Q_t is calculated by computing the weighted sum of the individual case relevance scores, weighting each by the similarity between Q_t and each Q_i . In this way, results which come from retrieved cases (C_1, \dots, C_n) whose query is very similar to the target query are given more weight than those who come from less similar queries. The relevance of a Result R_j to a target query Q_t and the case library comprising of cases from C_1 to C_n cases is expressed as:

$$\frac{\sum_i Relevance(R_j, C_i) * Similarity(Q_t, C_i)}{\sum_i Exists(R_j, C_i) * Similarity(Q_t, C_i)} \quad (1)$$

Similarity between the query and case is computed by finding the similarity between the query and case queries. We are using Jaccard Similarity as the similarity metric in our design. In this way, for given user, with query Q_t we produce a ranked list of results R_j that come from the community's case base and that, as such, reflects the past selection patterns of this community. If the case base doesn't retrieve cases or the similarity confidence of the retrieved results is less than a user specified threshold t then, Q_t is used by the meta-search module to retrieve a set of web search results.

The top results obtained either from the case base or the meta search engines (when retrieved results similarity to the case problem is below a minimum threshold) are shown to the user. In this way, results that have been previously preferred by community members are either promoted or marked as relevant to provide community members with more immediate access to results that are likely to be relevant to their particular needs. This framework promotes community preferred results and conversations to the user.

Discussion

Figure 2 displays the Cobot research prototype³. There are many technical challenges in community based information and recommendation systems. Instead of relying on search engines that inundate the user with a multitude of information, Cobot models the information finding task as a collaborative interaction process. The user describes his need in natural language which is modeled via text conversations

³www.cobothhealth.com



Figure 2: Cobot Interface

familiar to most users. In the Language understanding stage, there are several technically challenging bottlenecks that exist to capture the knowledge that is generally present in some external web repository. Improving the coverage and reliability of syntactic analysis, semantic parsing and extraction in conversations with near real time delivery of results to keep the users engaged requires highly sophisticated and robust taggers, parsers and classifiers. We are developing modules that identify simple factoid questions in conversations and directly intersperse high confidence answers in conversations.

Modeling users via their conversations also has several challenges. We attempt to construct models of human behavior to make the machine directly infer the knowledge and information need of a user so as to provide him with interesting recommendations. We are building ‘interestingness’ heuristics to capture knowledge for diversity and exploration of the topics that the user may be interested in. We are building spreading activation based topic exploration on the user model augmented with UMLS ontology hierarchy exploration to not only learn new concepts in the user model but also unlearn the knowledge that the user may longer be not interested in.

One problem we face in our case based reasoner is that when this module is triggered, we sometimes get recommendations that are not as good as the recommendations that come from web search. Our case based reasoning module is much faster than the web search and recommendation generation module. With usage, query-recommendation episodes grow in size quickly with case based maintenance becoming an important issue. Since the real time web is changing very rapidly, we randomly trigger web search on our backend to bring in new episodes in the case base. One of the features we are working on is to allow users to directly edit Cobot

recommendations in the conversations to create persistent user generated knowledge mixed with useful recommendations from external sources.

References

- Allen, J.; Chambers, N.; Ferguson, G.; Galescu, L.; Jung, H.; and Taysom, W. 2007. Plow: A collaborative task learning agent. In *In Proc. Conference on Artificial Intelligence (AAAI)*, 22–26. Springer-Verlag.
- Aronson, A. R. 2001. Effective mapping of biomedical text to the umls metathesaurus: The metamap program.
- Chen, L., and Sycara, K. 1998. Webmate: A personal agent for browsing and searching. In *In Proceedings of the Second International Conference on Autonomous Agents*, 132–139. ACM Press.
- Cross, R.; Rice, R. E.; and Parker, A. 2001. Information seeking in social context: structural influences and receipt of information benefits. *IEEE Transactions on Systems, Man, and Cybernetics, Part C* 31(4):438–448.
- E. Buyko, J. Wermter, M. P. U. H. 2006. Automatically adapting an nlp core engine to the biology domain. In *In Proceedings of the Joint BioLINK-Bio-Ontologies Meeting, Fortaleza, Brasil*, 65–68.
- Evans, B. M., and Chi, E. H. 2008. Towards a model of understanding social search. In *CSCW '08: Proceedings of the ACM 2008 conference on Computer supported cooperative work*, 485–494. New York, NY, USA: ACM.
- Fox, E. A.; Hix, D.; Nowell, L. T.; Brueni, D. J.; Rao, D.; Wake, W. C.; and Heath, L. S. 1993. Users, user interfaces, and objects: Envision, a digital library. *J. Am. Soc. Inf. Sci.* 44(8):480–491.
- Horowitz, D., and Kamvar, S. D. 2010. The anatomy of a large-scale social search engine. In *WWW*.
- Karasavvidis, I. 2002. Distributed Cognition and Educational Practice. *Journal of Interactive Learning Research* 11–29.
- McCarthy, K.; McGinty, L.; Smyth, B.; and Salamó, M. 2006. The needs of the many: A case-based group recommender system. *Advances in Case-Based Reasoning* 4106:196–210.
- McInnes, B.; Pedersen, T.; and Pakhomov, S. 2009. UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*.
- Pazzani, M. J., and Billsus, D. 2007. Content-based recommendation systems. *The adaptive web: methods and strategies of web personalization* 325–341.
- Pradhan, S.; Ward, W.; Hacıoglu, K.; and Martin, J. H. 2004. Shallow semantic parsing using support vector machines.
- Sahay, S., and Ram, A. 2010. Conversational framework for web search and recommendations. In *In Proceedings of the ICCBR Workshop on Reasoning from Experiences on the Web*.
- Sahay, S.; Mukherjee, S.; Agichtein, E.; Garcia, E. V.; Navathe, S. B.; and Ram, A. 2008. Discovering semantic biomedical relations utilizing the web. *TKDD* 2(1).
- Smyth, B.; Briggs, P.; Coyle, M.; and O’Mahony, M. P. 2009. A case-based perspective on social web search. In *Proceedings of the 8th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development*, 494–508. Berlin, Heidelberg: Springer-Verlag.
- Ybarra, O.; Burnstein, E.; Winkelman, P.; Keller, M. C.; Manis, M.; Chan, E.; and Rodriguez, J. 2008. Mental Exercising Through Simple Socializing: Social Interaction Promotes General Cognitive Functioning. *Pers Soc Psychol Bull* 34(2):248–259.