Supporting Personalized Information Exploration through Subjective Expert-created Semantic Attributes

Cormac Hampson, Owen Conlan
Knowledge and Data Engineering Group
Trinity College Dublin
Ireland
{hampsonc, owen.conlan}@cs.tcd.ie

Abstract—**Ordinary users are finding it increasingly difficult to explore the large volumes of diverse data they encounter in their everyday lives. Techniques based on data mining algorithms are useful but they tend to be too complex for casual users to work with effectively. Furthermore, these techniques don't allow the user to engage with the information using semantics meaningful to them. Semantically enriched and personalized data exploration is seen as an essential step to support such users. Moreover, by allowing these users to leverage and personalize the subjective insights and knowledge of experts, more relevant and useful information can be discovered and interesting correlations drawn. In order to support these domain specific explorations, a prototype architecture named SARA (Semantic Attribute Reconciliation Architecture) has been built, and its underlying methodology, implementation and initial evaluation are described within this paper.**

Keywords-Semantic Attributes; Subjectivity; Personalization; Information Exploration; Domain Experts

I. INTRODUCTION

In recent years, the sheer volume of digitized content and its exponential growth in all domains has necessitated better techniques to allow users to correlate information, and elicit useful knowledge from separate data sources. As such KDDM (Knowledge Discovery and Data Mining) techniques [1] have increasingly found a niche in both research and commercial environments, enabling interesting relationships and trends to be revealed within large data collections, despite these attributes not being explicitly encoded. However, the complexity of these approaches has meant it is difficult for end-users to explore the underlying data through meaningful semantics. Moreover, users are usually not willing to devote time and resources toward formal methods of knowledge seeking [1]. In many of the KDDM approaches used (e.g. statistical based, case based, neural networks and probability approaches) the importance of the user, specifically their expertise, motivation and goals, has often been overlooked. This has led some KDDM practitioners to assert that a pure focus on data, data structures and machine learning algorithms under-emphasizes the complex role of the human user [2].

In contrast, the main aim of human-centered computing is to satisfy the user by allowing them to make queries and see responses in their own terminology [3]. In effect, this means translating easily computable low-level data into high-level concepts or terms that are intuitive to users. Furthermore, systems should understand the semantics of a query and not just the underlying computational features. This practice is known as bridging the semantic gap [4].

The emerging field of HCIR (Human Computer Information Retrieval) aims to empower people to explore large-scale information bases, but demands that people also take responsibility for this control by expending cognitive and physical energy [5]. Marchioni devised a list of goals for any system that enables the user more control in determining relevant results in a search. The importance of the user (supported by information professionals) in this exploratory process is emphasized by the following inclusions [5]:

- Systems should aim to get people closer to the information they need, no longer only delivering the relevant documents but also providing facilities for making meaning with the document's contents.
- Systems should increase user responsibility as well as control; that is, requiring human intellectual effort, and rewarding it.
- Systems should have flexible architectures so they may evolve to increasingly more demanding and knowledgeable users over time.
- Systems should support tuning by end users and especially by information professionals who add value to information resources.

As a result, this research will take these goals and specifically address the need for a framework that supports a user's exploration of separate information sources, by enabling users to leverage expertise about the domain. Specifically this framework will provide operations to client applications that enable the personalization of a user's requests. This personalization will be in terms of the subjective insights encoded by domain experts into semantically meaningful concepts. By making this expert knowledge accessible to users within the scope of their search, it will enable them to tailor the semantics to their current preferences, and further support their exploration of large data repositories. Moreover, these hooks into a domain will support any applications using the framework to create

complex and semantically meaningful requests, and to present the results in a highly visual way to end-users.

As the prevalence of and dependence on electronic data escalates, and the increasing use of such repositories by casual users, the need for user-friendly systems capable of supporting meaningful exploration will be greatly increased. This paper will investigate the challenges of designing such a framework and its accompanying methodology. Section 2 describes some work related to this research with section 3 detailing the underlying design, and the semantic attributes which are central to the framework's operation. In section 4 there is a description of a use case, as well as information on the prototype implementation and evaluation of the architecture. Section 5 summarizes the paper and details the future direction of this research.

II. Related Work

The main aim of this research is to support, in a personalized fashion, user exploration of information stored over separate data sources. This section briefly discusses some pertinent work related to this field.

There are an increasing number of web based applications such as Google Base [6] and FreeBase [7], which allow complex queries to be made over large data repositories. Google Base enables users to upload structured or unstructured content and for this data to be labeled with predefined or custom attributes. Other users can then search this repository using free text or by making selections from fields and categories. Although it does allow relatively complex queries to be formed (e.g. find any multimedia jobs advertised within 30 miles of London that pays more than 50k salary), its main focus is on helping users to find relevant items, listings, and events, and it is unable to reconcile similar topics across information categories.

FreeBase [7] is an online knowledge base containing structured data harvested from many different sources. It allows free text queries to be run over its content, but its real strength over conventional search engines, such as Google and Yahoo!, comes in allowing users to create complex queries via its MQL query language. Freebase has a freely accessible API which allows external applications to leverage its knowledge base by sending queries to it remotely. However, there is still no easy way for users unfamiliar with MQL to create the complex queries in an easy or intuitive fashion, nor is it possible to customize the FreeBase framework to work with different information repositories.

One research project in development that aims to tackle the issue of allowing complex queries to be reconciled over multiple domains and services is NGS (New Generation Search) [8, 9]. There are an increasing amount of search services on the web, however they typically work in isolation covering single domains. The NGS framework aims to handle queries that span many domains such as "find all database conferences held within six months, in locations whose seasonal average temperature is 28°C, and for which a cheap travel solution exists". Thus, one of its main challenges is to combine the results of these separate services in an intelligent way. A limitation of this framework is that it doesn't allow users to incorporate expert knowledge into their queries to support deeper exploration of the sources. Moreover, this research project is still in a relatively early stage, and it does not specify yet how its generic framework will enable such specific queries to be formed easily by the user.

The previous examples can be seen as broad horizontal search platforms; however there is increasing interest in domain specific search for specialized areas or large enterprises. This is because generic search technology works well for general search but starts to degrade when addressing the needs of users in particular fields such as healthcare and finance [10]. In terms of increasing semantics in domain specific search, MedStory [11] is one of the leading exponents. Medstory aims to allow a user to search complex fields on the web intelligently and has started with the health and medicine domain to demonstrate the technical capabilities of its framework. The aim of MedStory is to synthesize the meaning of every user search in terms of health and medicine. Then through its interface it allows users to refine their initial queries to help locate the information most relevant to their question. Currently this is working in the single domain of health and medicine, but it purposes to be a generic framework for domain-specific search. Unlike Freebase or GoogleBase, Medstory will not provide specific answers to queries. Instead it presents a number of documents that it thinks the answer will lie within. Hence, Medstory will not allow queries to be reconciled across multiple information sources from a domain.

While not suitable for general Web searches, an expert system can help users find reliable information in a narrow area such as medicine or accounting [10]. Thus, many tools have been designed in order to capture domain knowledge from experts. However typically these tools [12] are very bespoke to their domain and designed to capture very specific knowledge. Thus, a generic platform for experts to encode their experience and knowledge of a domain would be very useful. This approach may not capture as much detail as a bespoke system would, however its generic nature would make its adoption and widespread implementation more likely.

Because the framework proposed in this paper can be seen as a type of mashup (combining data from one or more sources into a single integrated tool) there can be lessons learned from applications such as Yahoo! Pipes [13] and IBMs Damia [14]. These tools allow different information feeds to be combined together graphically to form a personalized information mash-up tailored to a user's preferences. It would be useful to employ a similar approach when supporting non-technical domain experts to encode their insights of a domain. Furthermore, the support for personalization in all the systems described is generally limited or non-existent, so incorporating such a facility into the proposed framework would be of benefit to end-users.

III. Design

This research addresses the need for a framework that supports a user's exploration of separate information sources

from a specific domain. The framework will provide specific operations to client applications, which will enable the personalization of a user's request in terms of the subjective insights encoded by domain experts. From the analysis of related work, the following high-level requirements were elicited:

1. Support semantically rich exploration across heterogeneous information sources from a domain.
2. Enable experts to encode their insights of a domain in a semantically meaningful way, in order to guide end users.
3. Facilitate user personalization of this subjective expertise, enabling them to build on expert opinion.
4. Support dynamism through the automatic updating of result sets from both the source and user sides.

This section discusses the design of the Semantic Attribute Reconciliation Architecture (SARA) and its underlying process model. Furthermore, it will detail the semantic attributes which are central to its operation, and outline its mechanism for consolidating data from multiple sources.

A. Process Model

In order to support a complete end-to-end solution, SARA has adapted the generic KDDM (Knowledge Discovery and Data Mining) process model specified by Kurgan and Musilek [1]. In their survey of the major process models in use by the KDDM community they specified a six step generic model which consolidates the information accumulated among these five major models. Because other disciplines like exploratory search encounter many of the same issues as KDDM, they can also benefit from such a consolidated process model. Steps 1 to 3 involve selecting, defining and preparing the raw data sources in terms of the domain, as well as the specification and encoding of domain expertise. There is already much research being conducted in the areas of data integration and automatic semantic mapping which may feed into these steps in future. SARA has its focus on personalized user exploration that leverages domain expertise, thus it was necessary to change step 4 of their generic process model (the selection of a specific data mining technique) to personalized exploration. Though incorporating all 6 steps of the process model, this fourth step is where the innovation in SARA is focused. Steps 5 and 6 of the process model, which evaluate the results and present them to the end user are outsourced by SARA to its client applications, and are supported through SARA's API. There is a lot of research being conducted in the area of information visualization and exploratory search that client applications will be able to incorporate. Figure 1 highlights how this six step process is applied to SARA.

B. Semantic Attributes

Semantic attributes are discrete encodings of domain expertise that can be combined together and personalized to support user exploration of an information domain. When an expert expresses an opinion on a dataset, it is inherently a more subjective way to register domain knowledge than the use of crowdsourcing or collaborative filtering techniques. Semantic attributes encapsulate these subjective insights of a domain and enables them to be personalized and linked according to an end-user's preferences. Though more

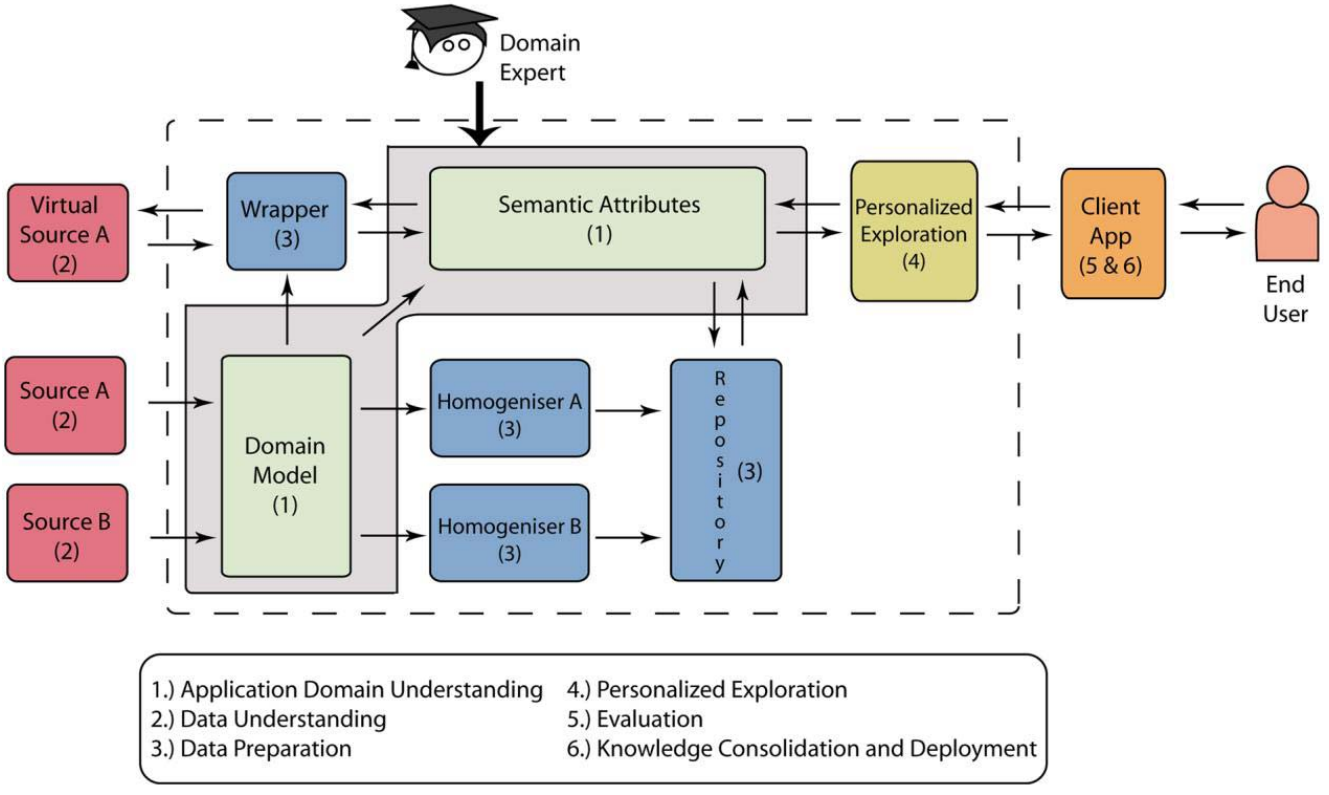

Figure 1. Generic KDDM process model as adapted and applied to SARA

objective knowledge can be encoded as a semantic attribute, there is also considerable value in allowing very individual and subjective analyses of a domain to be leveraged and shaped by users.

Through personalizing the semantic attributes, users can control and combine them to create semantically meaningful and expertly informed queries across heterogeneous information sources. In turn, this helps to bridge the semantic gap as it supports users to explore low-level data through semantics that are meaningful to them. Each semantic attribute is encoded in a model which is then imported into SARA. This model contains some or all of the following:

- A semantically meaningful concept
- Links to domain metadata
- Parameters with rules defined by experts
- Personalizable elements
- Links to global functions

Semantic attributes are agnostic to what technology the expert defined rules are encoded in, with the current implementation using XQuery.

Typically semantic attributes describe distinct and indicative concepts from a domain. For instance in the music domain, semantic attributes can encompass characteristics such as recentness, popularity, duration and similarity. There are two main types of semantic attributes, Quantized and Instance. Quantized semantic attributes can divide a concept into a number of different parameters such as high popularity, average popularity and low popularity. This parameterization is quite subjective and reflects the expert's or end-user's (through personalization) perspective. In contrast, Instance based semantic attributes are not parameterized and typically try to find specific instances within a semantic attribute, such as Jazz or Folk in a music genre semantic attribute.

Semantic attributes can be quite broad and are inherently subjective, e.g. declaring the notion of a long song in terms of a music collection as a whole. Likewise, they can have a finer granularity referring to long songs in terms of artists or genres. Thus by employing both kinds of semantic attributes it is possible to distinguish tracks that are long in terms of a specific artist or genre (long pop songs might be considered short classical or jazz compositions), from those that are long in terms of the entire collection of songs. Hence, a powerful aspect of semantic attributes is that just a single one can be used to automatically calculate what songs have long duration or are early in their career, for thousands of different artists.

Semantic attributes can also be composed and balanced together by the user to form more complex queries tailored specifically to their needs. These are called semantic attribute queries, with the personalizable element being vital in enabling users to specify, if they wish to, what their interpretation is of a high quality audio file or a popular song etc. For instance, a sound engineer in the context of his studio work may define a high quality audio file as one that is uncompressed raw audio, whereas in the context of his home listening the same semantic attribute could be tweaked by him to include compressed MP3 files above a minimum value bitrate. Each semantic attribute will also include default values defined by the domain expert that allow informed queries to be run quickly without personalization.

Because semantic attributes are hand-crafted by experts and not just automatically extracted from datasets it is important to allow them to be created by non-technical experts in minutes. Hence an authoring tool with a wizard GUI has been developed to directly support this. This means that SARA gets the benefit of accurate human-created semantic attributes without the cost of large amounts of manual effort. Furthermore, because SARA allows new semantic attributes to be added seamlessly at any time by different users, it facilitates teams of experts in performing collaborative work, or to have different types of expertise exposed across the same domain. This diversity of expert perspectives encoded as semantic attributes empowers end-users to pick and choose the semantic attributes that are best suited to their needs. In some ways this is analogous to choosing a specific critic for guidance in a domain you have interest in, but are not an expert in. Thus their subjective critiques on movies, sport, politics, finance etc. can be appropriated by the end-user to help their exploration, but more importantly they can be tailored to greater match individual preferences.

C. Reconciliation of Sources and Dynamism

In order to support consolidated queries across sources, it is necessary to index inputted data to one or more superclasses and use these to bridge the individual sources. For instance, the music domain model could associate the trackDuration, chartPosition and trackBitrate metatags with the superclass Song and could associate artistName, concertScheduled, nationality, with the superclass Artist. When creating a semantic attribute, the type of superclass it will return must be specified. By declaring this, it allows complex semantic attribute queries (those containing individual semantic attributes associated to different superclasses) to be reconciled together automatically. To illustrate with a basic example, this would enable a user, in just one consolidated request, to find all the artists who are currently touring (a semantic attribute based on the Artist superclass), despite not having a top 40 hit in more than ten years (semantic attributes based on the Song superclass).

After a user makes a request of SARA, it gets broken down to individual semantic attributes and a set of results for each individual semantic attribute created. Then using set operators the results of each individual semantic attribute set get consolidated into a final result set. As specified in the previous example, if the semantic attributes are based on different superclasses, an automatic conversion process takes place so that the individual sets can interact.

The advantage of using sets in this way is that SARA can then support both user-driven and source-driven dynamism. On the user side, if someone is exploring the data sources and updating their query dynamically (perhaps through a visual interface), the relevant semantic attributes can have their bounds adjusted accordingly and the result set populated to reflect the updated status of the information. Likewise from the other side, if there are dynamic sources

connected to SARA and new information is received, this data can be checked against the current semantic attribute filters. If this new information falls within the boundaries of one of the individual sets, this part of the query can be rerun to populate the final result set appropriately.

IV. IMPLEMENATION,USE CASE AND EVALUATION

Based on the design considerations outlined in the previous section, a prototype implementation of SARA was undertaken. As an initial use case, the generic framework implemented was tailored to the personal music domain. However any domain with suitably rich content sources could have been chosen and integrated into SARA in an identical fashion.

A. Design Time Implementation

Implemented in Java, SARA had three heterogeneous sources related to the music domain connected to it. These sources were a collection of iTunes music databases, a HTML archive containing UK music chart information, and various data harvested from web services exposed by Last.fm [15], an online radio station and social network. This selection of sources was part of the *data understanding* process described in the process model used by SARA. The next step in the use case was for a canonical domain model to be constructed by an expert in the area of interest. This is a key part of the *application domain understanding* process. In order to add an information source to the system, a bespoke homogenizer had to be created which normalized each source semantically, syntactically and structurally with reference to the domain model. This is part of the *data preparation* step in the process model. In future, semantic wrappers will be used to transform data from distributed sources to the domain vocabulary dynamically.

It is the domain expert who decides what key semantic attributes for the domain to encode into the system. Thus, the key challenge for them is to translate metadata from the domain model into useful characteristics that an end user could manipulate and combine. A simple example of this is the semantic attribute *AudioFile Quality* mapping onto the `<AudioFileType>`, `<TrackBitRate>` and `<TrackSampleRate>` metatags from the domain model. Each semantic attribute is encoded using the prototype authoring tool mentioned in the previous section. This tool aims to support domain experts, without computer programming experience, to create the semantic attributes and their associated rules through its GUI. In this implementation the rules which define the default and personalizable properties were encoded in XQuery and only required the WHERE part of a FLWOR expression [16]. Thus a default rule for a *high quality audio file* could be encoded as simply as `AudioFileType = 'MP3' and TrackBitRate >= 128 and TrackSampleRate >= 44,100.`

Once a semantic attribute is created in the GUI, it is imported into SARA where it is converted into a corresponding Java method. This method is instantly made available through SARA's API so that client applications that want to provide this particular semantic attribute to their users can access it immediately.

B. Runtime Implementation

Any client application connected to SARA can support the user to define their search in any way they want as long as it conforms to the SARA API. Hence, this can range from highly visual and interactive interfaces to more standard GUIs or query wizards. An initial demonstration application was created for this implementation which allowed users to select, combine and personalize the semantic attributes they wished to send to SARA, and then have the results displayed to them in XML. More visual demonstrator clients supporting exploration of photographs, movie information and academic publications are being developed at the moment.

Typically user defined queries combined a number of personalized semantic attributes with the superclass they wanted to return (as mentioned earlier, in this use case it was either a *Song*, an *Album* or an *Artist*) which were sent to SARA via its API. When each semantic attribute query was received by SARA it was decomposed into its constituent semantic attributes. Each associated Java method was called individually and the appropriate parameters passed to it. Each semantic attribute returned an individual result set which could then be consolidated to give an overall result for that query. This result was sent back as XML to be rendered in the client application.

C. Evaluation

In order to perform an initial evaluation of the prototype architecture, an experiment was devised to show how users could create semantically meaningful requests in order to explore a domain of interest. The difficulties in doing quantitative evaluation of exploratory search systems, in comparison to the standard precision and recall metrics available in IR systems are well documented [18, 19]. Hence, this evaluation was focused on getting some initial qualitative feedback on SARA.

As mentioned previously, the data sources used in the evaluation were three separate iTunes databases totaling 7,000 songs, a UK music chart archive from 1994-2008, and information harvested from web services offered by Last.fm. Initially the domain expert selected eleven semantic attributes for the personal music domain and encoded them as XML. In accordance with our high-level requirements, this domain expertise could then be leveraged during the experiment. Twelve users (eight with strong IT skills and four who did not) participated in the experiment and ten complex tasks were demonstrated. This list included tasks such as *locating any songs in their collection that reached number 1 in the UK charts since 1994 and recommending five songs similar to these hits*. This particular task highlighted how users could create one

consolidated query that previously would have involved the correlation of information from three separate data sources. Other exploration tasks for this experiment included finding *any artists in the collection that had concerts currently scheduled despite not having had a hit since 1994* and locating *any long songs in the collection by artists similar in style to the Beach Boys*. The participants were encouraged to adjust these requests to take into account their definition of a *long* song, or a user's preferred time span or perception of a *hit* (top ten, top forty etc.). This highlighted how personalization and user driven dynamism (high-level requirements for the system) were facilitated by SARA. Users were also able to construct their own search queries.

Despite the relatively small sample set of users, it was possible to get indicators from them as to what was successful about the current implementation of SARA. User comments consistently stated that SARA supported them in creating complex requests in a trivial fashion and that correlations could be drawn between the results that would not have been possible otherwise. For instance, it was noted that it was now viable to group together all the artists in the collection that had not had chart success recently. Likewise comparisons could be drawn between online fans and the general charts, in how particular songs were rated. Furthermore, users described how they felt that they had sufficient control in personalizing the subjective semantics of their searches so that the information returned was relevant to their interpretation of the semantics. This facilitated them finding "a needle in a haystack". Even with a small number of semantic attributes available in this experiment, user's believed overwhelmingly that these provided a good initial starting point to explore the domain.

V. Future Work and Summary

The next phase of SARA's development will examine the use of encoding semantic attributes in SPARQL to support queries that run over RDF and Linked Data repositories. Likewise, further implementation of semantic mediators will negate the need for a central data repository and a priori homogenization of source data. As the number of semantic attributes for a domain grows there will be a need for further support within SARA for user modeling, so that appropriate semantic attributes can be presented to users according to implicit and explicit preferences. In parallel to this development, the prototype authoring tool for creating semantic attributes will be developed more so that it is even easier and more intuitive for non-technical experts to create new instances. Finally, there are a number of separate client applications in development that utilize SARA to support end users exploration of different domains in a highly visual fashion.

This paper has introduced SARA and its support for users who want to explore heterogeneous sources from a domain in a consolidated manner. How domain expertise can be encoded as semantic attributes was described, as well as their ability to be personalized to an end users preferences. The initial evaluation of the system proved positive, and its findings will be fed into the development of the next version of SARA.

VI. Acknowledgements

This research has been supported by The Irish Research Council for Science, Engineering and Technology: funded by the National Development Plan.

VII. References

[1] L. A. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining process models," The Knowledge Engineering Review, vol. 21, pp. 1-24, 2006.

[2] A. Nayak, "User centered approach for the design of knowledge discovery: systems used in technology management," in Portland International Conference on Management of Engineering and Technology. PICMET '99., 1999, p. 111.

[3] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," ACM Trans. Multimedia Comput. Commun. Appl., vol. 2, pp. 1-19, 2006.

[4] R. Zhao and W. I. Grosky, "Narrowing the semantic gap-improved text-based web document retrieval using visual features," IEEE Transactions on Multimedia, vol. 4, pp. 189-200, 2002.

[5] G. Marchionini, "Toward human-computer information retrieval," Bulletin of the American Society for Information Science and Technology, vol. 32, 2006.

[6] "Google Base. Post it on Base. Find it on Google", [Online]. Available: http://base.google.com. [Accessed April 23rd, 2009]

[7] "Freebase: A social database about things you know and love", [Online]. Available: http://www.freebase.com [Accessed April 23rd, 2009]

[8] D. Braga, D. Calvanese, A. Campi, S. Ceri, F. Daniel, D. Martinenghi, P. Merialdo, and R. Torlone, "NGS: a framework for multi-domain query answering," in IEEE 24th International Conference on Data Engineering Workshop, 2008. ICDEW 2008. , 2008, pp. 254-261.

[9] D. Braga, S. Ceri, F. Daniel, and D. Martinenghi, "Optimization of multi-domain queries on the web," Proceedings of the VLDB Endowment archive, vol. 1, pp. 562-573, 2008.

[10] S. J. Vaughan-Nichols, "Researchers MakeWeb Searches More Intelligent," Computer, vol. 39, pp. 16-18, 2006.

[11] "Intelligent search for Health & Medicine.", [Online]. Available: http://www.medstory.com/. [Accessed April 22nd, 2009]

[12] C. Duchêne, M. Dadou, and A. Ruas, "Helping the Capture and Analysis of Expert Knowledge to support Generalisation" The 8th ICA Workshop on Gernalisation and Multiple Representaion, A Coruna, Spain 2005.

[13] J. C. Fagan, "Mashing up Multiple Web Feeds Using Yahoo! Pipes," Computers in Libraries, vol. 27, p. 8, 2007.

[14] M. Altinel, P. Brown, S. Cline, R. Kartha, E. Louie, V. Markl, L. Mau, Y. H. Ng, D. Simmen, and A. Singh, "Damia: a data mashup fabric for intranet applications," 2007, pp. 1370-1373.

[15] "Last.fm Website, The Social Music Revolution", [Online]. Available: http:www.last.fm. [Accessed April 25th, 2009]

[16] "XQuery 1.0: An XML Query Language", [Online]. Available: http://www.w3.org/TR/xquery/#id-flwor-expressions. [Accessed April 17th, 2009]

[17] R. W. White, G. Muresan, and G. Marchionini, "Report on ACM SIGIR 2006 workshop on evaluating exploratory search systems," 2006, pp. 52-60.

[18] Y. Qu and G. W. Furnas, "Model-driven formative evaluation of exploratory search: A study under a sensemaking framework," Information Processing and Management, vol. 44, pp. 534-555, 2008.