ESpace
Web-scale Integration One Step at a Time

Kajal Claypool, Jeremy Mineweaser, Dan Van Hook, Michael Scarito
Massachussetts Institute of Technology/Lincoln Laboratory, 244 Wood Street, Lexington, MA, U.S.A.

Elke Rundensteiner
Computer Science Department, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA, U.S.A.

Keywords: Data integration social bookmarking pay-as-you-go web-integration.

Abstract: In this paper, we take the position that a flexible and agile integration infrastructure that harmoniously and transparently oscillates between and supports different levels of integration – *loose* or *partial* integration on one end of the spectrum and *tight* or *full* integration on the other end of the spectrum – is essential for achieving large Web scale integration. Furthermore, domain knowledge provided by users/domain experts is essential for improving the quality of integration between resources. We posit Web 2.0 or "social Web" technologies, can be brought to bear to facilitate implicit user-driven, web-scale integration at different levels. In this paper, we present *ESpace*, a prototype for a pay-as-you-go integration framework that supports loosely to tightly integrated resources within the same infrastructure, where loose integration is supported in the sense of pulling resources on the web together, based on the tag meta-information associated with them, and tight integration is a representation of classic schema-matching based integration techniques. This is but the first step in enabling web-scale pay-as-you-go integration by providing fine-grained analysis and integrating substructures within resources – achieving tighter integration for select resources on the user's behest.

1 INTRODUCTION

There is a gradual but growing paradigm shift in the way users operate on the World Wide Web (Web). Web users today are not satisfied by merely viewing published content on the Web but want to actively engage in the larger Web community and share their knowledge. This paradigm shift has resulted in an ever-growing repository of published and shared Web-accessible data. They, the Web users, are interested in chronicling their opinions (*blogging*), adding to the knowledge pool on different subject matters (*wikis*), sharing what they have read (*social bookmarking*) and are interested in the potential to connect to like-minded people within their community or across the globe, either explicitly (*social networking*) or implicitly.

The challenge is in being able to discover and extract the right "nuggets" of knowledge from the resulting collective pool of "community intelligence" on an as needed basis, where the nuggets range from finding users with similar interests, finding documents that are similar to those already read by a user, to discovering topically linked communities within the larger community. The challenge in finding the right "nuggets" of knowledge is in formulating links between different data based on the available information, that is in integrating the data on the Web scale.

While traditional data integration techniques (Baker et al., 1998; Doan et al., 2001; Halevy et al., 2003; Wiederhold, 1992; Bright et al., 1994; Bergamaschi et al., 2001; Berlin and Motro, 2001; Haas et al., 1999; Madhavan et al., 2001; Do and Rahm, 2002) can be brought to bear to achieve this, they in of themselves are not agile nor flexible enough to address this large scale Web integration. Traditional data integration techniques are grounded in schemas, requiring design-time schema level matches to drive run-time integration of data. While some Web sources lend themselves to this mode of integration, there are many others that do not even have a well-established schema. Moreover, apriori establishing schema level matches between many of the resources may well be a colossal waste of time and resources, as there may never be a user desire to search the resources in an integrated manner.

Claypool K., Mineweaser J., Van Hook D., Scarito M. and Rundensteiner E. (2009).
ESpace - Web-scale Integration One Step at a Time.
In *Proceedings of the 11th International Conference on Enterprise Information Systems - Databases and Information Systems Integration*, pages 247-252
DOI: 10.5220/0002161802470252

In this large Web scale integration it is, thus, essential to establish an infrastructure that allows for *loose* or *partial* integration on one end of the spectrum, but that is also able to transition to *tight* or *full* integration on the other end of the spectrum harnessing available information, such as the schema, while fulfilling user needs. While clearly richer, finer nuggets of information can be extracted from the more tightly integrated resources, loosely integrated resources have the advantage of delivering partial nuggets of information at very little cost. Figure 1 highlights the model for the *pay-as-you-go* (Maier et al., 2005; Doan, 2008; Sarma et al., 2008) integration wherein, resources in the system co-exist in various integration stages, from no integration to partial integration to full integration. The richness of the nuggets, indicated by the semantic interoperability, increases with the level of integration and is directly proportional to the effort required in establishing the integration links between the resources.

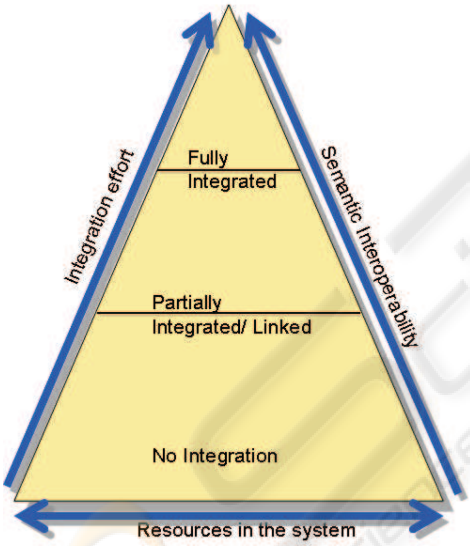

Figure 1: Supporting Different Levels of Integration.

We posit that the popularity and the collaboration potential unleashed by today's wildly read-write Web, often referred to as Web 2.0 or the "social Web", can be utilized and brought to bear to facilitate implicit user-driven, web-scale integration, and push towards a flexible integration infrastructure that can harmoniously and transparently support different levels of integrated systems.

In this paper, we present the vision of such a flexible integration system: *ESpace*. *ESpace* is a scalable, pay-as-you-go integration system that provides a *community collaboration infrastructure*, together with a suite of algorithms that mine the tagging activities of users to establish integration links between resources as well as users. *ESpace* harnesses the collective intelligence of many to help individuals as well as whole communities by: (1) collecting and consolidating resources on the web along with tags annotating the resources in a sharable and flexible community graph model; (2) designing and realizing a comprehensive set of services that mine this collective intelligence in the form of tagging activities of users to extract rich integration relationships among users and artifacts within the community graph model; (3) exploiting this so enriched community graph model to provide value-added services to the community and its individuals, ranging from recommendations of resources, to the computation of evolutionary trends and shifts of expertise, users or topics within the community, as well as to assisting others to locate subject experts or like-minded users to gain strong confidence into the identified resource.

The rest of the paper is organized as follows. Section 2 gives a high level overview *ESpace*. Section 3 details the community graph model, while Section 4 outlines some of the algorithms that are currently being implemented to harness tagging activities for integration links. Section 5 highlights some the benefits that can be realized via the *ESpace* loose integration infrastructure. We conclude in Section 6.

2 OVERVIEW

Figure 2 highlights the key concepts of the *ESpace community integration* architecture. *ESpace* is presented to its users primarily as a social bookmarking site enhanced and enriched via a set of services that deliver knowledge mined from the collective intelligence of the community in various forms. As a social bookmarking site, *ESpace* provides bookmarklets that allow users to bookmark and tag web documents, and consequently add to the collective graph that represents the community information base. In general, to accomodate a large variety of users, *ESpace* aims to support a range of input tools, such as integrating with browser bookmarking tools, providing ingest of bibtex files, and enabling both a simple one click and a more full-fledged bookmarklet for tagging documents. Information provided by the user, in the form of the *user*, the bookmarked *document*, the *tags* that represent the user's take on the document, together with annotations such as the time of creation or of last update, is captured as a graph, and subsequently merged with the collective community graph stored in the *Common Logical Repository*.

The intellectual merit of this work lies in the

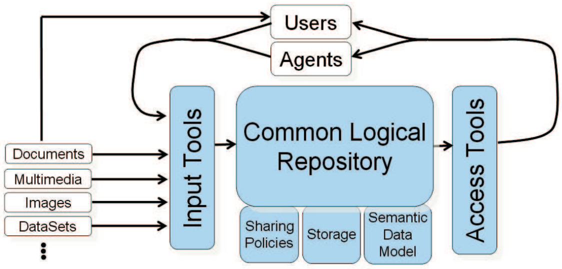

Figure 2: Architectural Overview of *ESpace*.

development of a set of *agents* that perform value-added services such as: *data cleanser* to, for example, ensure that mis-spelled tags are corrected, synonymous tags are identified correctly, and same documents identified by two different URLs [1] are merged; *artifact linker* to, for example, infer and establish relationships between two users, two documents or in some cases even two tags, based on the collective information garnered by the community as a whole and the individual user; *top-k recommender* to, for example, recommend topics, users ("friends"), documents, and in some cases even tags of potential interest to the community as a whole, or to provide personalized recommendations of the same to a user of the system; and *trend analyzer* to provide a birds-eye view of the collective knowledge of the community or an individual user over time delivering, for example, the evolution of the topics of interest or the user community as well as providing insights into the general trends in the knowledge and behaviors of users over some time period.

Much like users of the system, each agent adds to the collective information by overlaying their derived knowledge over the community graph as a whole or a subset of the graph when appropriate, providing continuous refinement of integration links over time. This overlay of information can be *materialized*, for example, relationships denoting similarities between users, wherein the relationships are physically added to the community graph, or *virtual*, for example, top "friends" or hot topics, wherein the information is computed on an as needed basis.

The usefulness of information collected by the user or derived by the agents is directly co-related to the tools that are provided to access and present the knowledge to the users. To accomodate a large variety of users *ESpace* provides several access tools such as classic keyword and strutured metadata searches; visualization tools such as Clustermap, Timeline, and Clouds (see Figure 3) to provide different views, ranging from "friends" of users to similar documents to user activity over time to "friends" over time, of the agent-augmented community graph; as well as visualization of recommendations, such as list based document recommendations, or user recommendations that also provide an opportunity for users to provide feedback for improved agent behaviors.

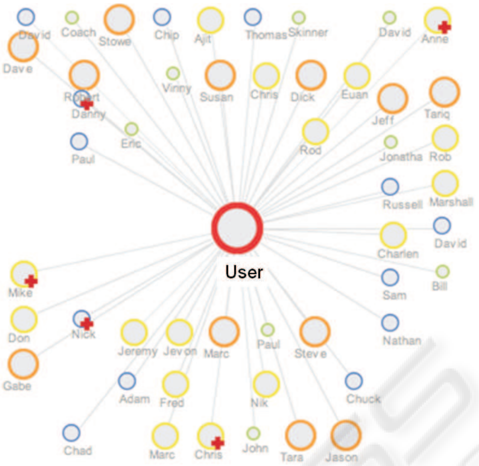

Figure 3: Clustermap Visualizing a User's Neighborhood.

3 *ESpace* COMMUNITY GRAPH MODEL

Social bookmarking sites and associated Web 2.0 technologies generally rely on relational databases and site-specific schema definitions to store the information shared by their users. While all social bookmarking systems store basic information on user, their bookmarked documents, and associated tags, each site adds its own variations: some may timestamp each document tagged by the user, while others may allow users to define hierarchical relationships between tags. Some recent work has looked at developing formal models for representing *folksonomies*, where a folksonomy represents the relationship between the user's documents and tags, and identifies the vocabulary set used by a user. While this is a needed first step, (1) it does not inherently treat users, documents and tags as first-class citizens; and (2) it is not flexible enough to express arbitrary, integration relationships between different artifacts[2] such as those defined by different agents; and (3) it is not extensible enough to capture information such as the strength of the relationship between two artifacts, indicating the degree of integration between two arti-

[1] For example, a paper posted on two different web sites.

[2] We refer to users, documents and tags as artifacts, when no distinction between them is required.

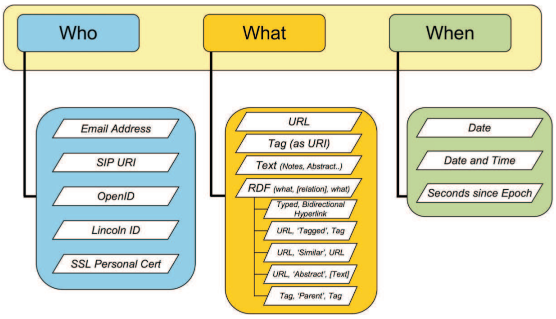

Figure 4: Community Graph Model.

facts, or distinguish between *explicit* relationships established by users and *implicit* relationships inferred by various agents.

To address these shortcomings, we provide an extensible and flexible *community graph model* where the nodes of the model represent the artifacts of our system, and the edges in the model represent implicit or explicit relationships between the different artifacts irrespective of whether the artifacts are users, documents or tags. Figure 4 gives a high-level overview of the community graph model as Resource Description Framework (RDF) triples of the form $<$ `who, what, where` $>$. The community graph model is a nested triple model that allows a range of relationships such as $<$ User$_i$, URL$_i$, timestamp $>$ and $<$ URL$_i$, `sameAs`, URL$_j$ $>$, to be expressed and enforced between the different entities in the model.

In addition to the community graph model, we provide a set of graph operators to facilitate functions such as merging a user-graph with the larger community-graph and extracting an artifact-centric graph from the larger community graph, in addition to core functions such as adding and removing artifacts or relationships to or from the graph.

4 ARTIFACT LINKERS: ESTABLISHING LOOSE INTEGRATION LINKS

The community graph model captures the relationship between the *user*, the *document* that is bookmarked, and the *tags* utilized by the user to annotate the document. These explicit relationships between artifacts represent the raw information that can in turn be exploited to infer integration/similarity relationships between the artifacts. For example, if two documents $\mathtt{d}_1$ and $\mathtt{d}_2$ are bookmarked by the same user $\mathtt{u}_1$ with exactly the same tags { $\mathtt{t}_1$, $\mathtt{t}_2$, and $\mathtt{t}_3$ }, it can be inferred that the documents provide similar knowledge for the user and thus conceptually may fall into the same category. Based on this knowledge, an *implicit similarity* relationship, termed *similarity* for brevity, between the documents $\mathtt{d}_1$ and $\mathtt{d}_2$ can be established. This similarity relationship represents a "loose" integration relationship between the two documents. This example can be further extended to establish similarity between documents that are tagged similarly by distinct users, or between users that share similar tag and document choices, or even between tags that are used equivalently. It should be noted that not all implicit similarity relationships are equal. For example, inferred similarity between documents $\mathtt{d}_1$ and $\mathtt{d}_2$ that are tagged with the same tags by the same user would be *stronger* than between documents that share only some of the tags. Moreover, the strength of the relationship can be refined further if it can be established that the documents were authored by the same person. This schema-level information transitions the relationship between the two documents from a "loose" integration to a "tight" integration relationship.

The *artifact linkers* infer implicit similarity relationships between two artifacts of the same type, irrespective of whether the artifacts represent users, documents, or tags; and determine the strength of the implicit similarity relationship between the artifacts. Academic and commercial research has broached this problem of inferring similarities from several angles. Collaborative filtering defines similarity of varying strengths between users based on how their ratings on items correlate (Herlocker et al., 2004; Bell et al., 2007; Konstan, 2004). Document similarity defines similarity of varying strengths between documents based on their content (Lee et al., 2005; Hammouda and Kamel, 2004; Paepcke et al., 2000). Semantic similarity defines similarity of varying strengths between words (or tags) based on the likeness of their meaning or semantic content. Traditional schema integration techniques (Baker et al., 1998; Doan et al., 2001; Halevy et al., 2003; Wiederhold, 1992; Bright et al., 1994; Bergamaschi et al., 2001; Berlin and Motro, 2001; Haas et al., 1999; Madhavan et al., 2001; Do and Rahm, 2002) establish similarity of varying strengths based on their schema. More recently in the context of social bookmarking, folksonomy similarity investigates how the folksonomies of different users can be correlated. Using these as building blocks, we now develop composite algorithms to support establishment of integration links between different artifacts.

To infer implicit similarity relationships between artifacts, we develop novel algorithms that harness the collective intelligence of the community graph or a subset to establish similarity between users, documents and tags. In particular, we provide three classes of algorithms, *user-user*, *document-document*, and *tag-tag*, that build upon and exploit implicit similarity relationships established by the other algorithms.

User-User. This class of algorithms determines the strength of the implicit similarity relationships between pairs of users. These include: (1) singleton set-intersection algorithms that determine the strength of the similarity based on overlap in documents and tags between the two users, along with algorithms that incorporate notions of signal to noise ratio in the documents and tags; and (2) composite algorithms that provide for meaningful combinations of the different singleton algorithms to determine an aggregated similarity strength. Additionally, the algorithms incorporate additional constraints such as the time of tagging, and knowledge such as already established document-document and tag-tag similarity relationships to age and fortify the strength of the similarity relationship, respectively.

Document-Document. This class of algorithms determines the strength of the implicit similarity relationship between pairs of documents. Similar to user-user algorithms, set-intersection based methods that determine similarity strengths based on document tags and document "owners", that is the users that bookmarked the documents, and combinations thereof have been developed. In addition, document similarity techniques (Lee et al., 2005; Hammouda and Kamel, 2004; Paepcke et al., 2000), and already established user-user and tag-tag similarities are used to augment the strength of the similarity relationship determined by the tag-based methods.

Tag-Tag. This class of algorithms determines the similarity relationship between pairs of tags. Set-intersection based methods that determine the co-occurence of tags based on overlap in documents and users, together with mapping of tags to common ontological concepts have been developed. Techniques from information retrieval such as semantic similarity together with the basic spell-checking and stemming algorithms are exploited to determine tag-tag similarity. Additionally, already established user-user and document-document similarities are used to augment the strength of tag-tag similarities.

5 BENEFITS OF LOOSE INTEGRATION

Social tagging systems capture a rich set of information. This information, the collective pool of knowledge for an individual user or the community as a whole, together with established loose integration links between artifacts can be leveraged to provide a wider variety of recommendations, from users to documents to tags, than have been possible in traditional recommendation systems (Amer-Yahia et al., 2008). The similarity relationships between different artifacts in *ESpace* can be leveraged to provide rich and diverse types of recommendations customized for individual members of the community. For instance, users may receive recommendations on *friends*, meaning other users in the community that share their interests; or on *documents* that are of potential interest to the individual based on the documents tagged by their friends and/or experts; or on *topics* that are of potential interest based on the topical areas favored for instance by their friends; or on *tags*, that is the tags that are of potential applicability to the document being bookmarked by the user. Last but not the least, users can also receive recommendations on *experts*, that is on other users that are singled out not only because they share the individual's interests, but who have also categorized a mass of literature in topic areas of interest. In all of these cases, interests of users are inferred by the documents they bookmark and the tags they use, that is the user's tag cloud and the integration links established between the different artifacts as core source of knowledge.

While recommendations target individual users, the *ESpace* information model as a whole can also be leveraged to provide trend analysis – a pulse of the community, a bird's eye view of both the current state of the community, as well as its evolution over time. In particular, we top-*k* analysis that identifies *communities of interest*, that is groups of users in the community that are clustered by their shared interests; *top-k users*, lead contributors to the communities of interest as categorized by their bookmarking activity; *hot topics*, the topical areas of interest most active within the community; *topical experts*, users singled out for categorizing a mass of literature within each hot topic; and last but not the least *top tags*, a measure of the most frequently used annotations in the community.

These two areas, recommendations and trend analysis, represent two rich areas of research that can be enabled via the loose integration model of *ESpace*, highlighting the benefits and value-added functionality that can be brought to its users.

6 CONCLUSIONS

ESpace represents a community collaboration infrastructure that harnesses the social bookmarking activities of its users to establish loose integration links between data sources as well as its users. While focusing on loose and partial integration, *ESpace* empowers its users to easily share information through a social bookmarking-inspired approach, while at the same time providing an enriched set of services that deliver knowledge mined from the collective intelligence of the community. The loose integration links between different artifacts lay the foundation for providing rich recommendation services, ranging from the *global* recommendations applicable for the entire community to *personalized* recommendations based on the context; as well as trend analysis strategies that provide a birds-eye view of the collective community knowledge as they change over time.

In this paper, we focus on loose integration in the sense of pulling resources on the web together, based on the tag meta-information asssociated with them. We posit that harnessing Web 2.0 user activities is an essential ingredient for achieving the holy grail of web-scale integration and *ESpace* is but a first step towards achieving a flexible integration infrastructure that supports loosly and tightly integrated resources.

REFERENCES

Amer-Yahia, S., Galland, A., Stoyanovich, J., and Yu, C. (2008). From del.icio.us to x.qui.site: recommendations in social tagging sites. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1323–1326, New York, NY, USA. ACM.

Baker, P., Brass, A., Bechhofer, S., Goble, C., Paton, N., and Stevens, R. (1998). TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources: An Overview. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, ISMB98*.

Bell, R., Koren, Y., and Volinsky, C. (2007). Modeling relationships at multiple scales to improve accuracy of large recommender systems. In Berkhin, P., Caruana, R., and Wu, X., editors, *KDD*, pages 95–104. ACM.

Bergamaschi, S., Castano, S., Vincini, M., and Beneventano, D. (2001). Semantic integration of heterogeneous information sources. *Data and Knowledge Engineering*, 36(3):215–249.

Berlin, J. and Motro, A. (2001). AutoPlex: Automated Discovery of Content for Virtual Databases. In *CoopIS*, pages 108–122.

Bright, M., Hurson, A., and Pakzad, S. H. (1994). Automated Resolution of Semantic Heterogeneity in Multidatabases. *TODS*, 19(2):212–253.

Do, H. and Rahm, E. (2002). COMA - A System for Flexible Combination of Schema Matching Approaches. In *vldb*.

Doan, A. (2008). Building structured web community portals via extraction, integration, and mass collaboration. In Ho, T. B. and Zhou, Z.-H., editors, *PRICAI*, volume 5351 of *Lecture Notes in Computer Science*, page 3. Springer.

Doan, A., Domingos, P., and Halevy, A. (2001). Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. In *sigmod*.

Haas, L., Miller, R., Niswonger, B., Roth, M., Schwarz, P., and Wimmers, E. (1999). Transforming Heterogeneous Data with Database Middleware: Beyond Integration. *IEEE Data Engineering Bulletin*, 22(1):31–36.

Halevy, A., Ives, Z., Mork, P., and Tatarinov, I. (2003). Piazza: Data management infrastructure for semantic web applications. In *World Wide Web Conf., 2003.*, pages 20–24.

Hammouda, K. M. and Kamel, M. S. (2004). Document similarity using a phrase indexing graph model. *Knowl. Inf. Syst.*, 6(6):710–727.

Herlocker, J. L., Konstan, J. A., Terveen, L., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53.

Konstan, J. A. (2004). Introduction to recommender systems: Algorithms and evaluation. *ACM Transactions on Information Systems*, 22(1):1–4.

Lee, M., Pincombe, B., and Welsh, M. (2005). An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259.

Madhavan, J., Bernstein, P., and Rahm, E. (2001). Generic Schema Matching with Cupid. In *vldb*, pages 49–58.

Maier, D., Halevy, A., and Franklin, M. (2005). From databases to dataspaces: A new abstraction for information management. *Sigmod Record*, 34(4):27–33.

Paepcke, A., Garcia-molina, H., Rodriguez-mula, G., and Cho, J. (2000). Beyond document similarity: Understanding value-based search and browsing technologies. *SIGMOD Record*, 29:2000.

Sarma, A. D., Dong, X., and Halevy, A. Y. (2008). Bootstrapping pay-as-you-go data integration systems. In Wang, J. T.-L., editor, *SIGMOD Conference*, pages 861–874. ACM.

Wiederhold, G. (1992). Mediators in the architecture of future information systems. *IEEE Computer*, 25(2):38–49.