Leveraging on Associations – a New Challenge for Digital Libraries

Martin Doerr
ICS-FORTH
Heraklion-Crete
Greece
+30 2810 391625
martin@ics.forth.gr

Carlo Meghini
CNR-ISTI
Via G. moruzzi, 1
Pisa, Italy
+39 050315 2893
carlo.meghini@isti.cnr.it

Nicolas Spyratos
University Paris-South
LRI – Bât 490
91405 Orsay Cedex, France
+33169156586
spyratos@lri.fr

ABSTRACT

Decades of research have been devoted to the goal of creating systems which integrate information into a global knowledge network. On the other side, Digital libraries have not overcome the traditional paradigm of delivering a document as ultimate objective. This paper argues that next-generation DL services must be built on accessing associations implicit or explicit in document collections and their metadata. It suggests a new approach to leverage associations based on (i) generic core ontologies of relationship and co-reference links (ii) semi-automatic maintenance of co-reference links by a new kind of service, and (iii) public engagement in the creation and development of the emerging association network.

1. INTRODUCTION

The Web has become an indispensable tool of modern culture. Powerful, but relatively crude search engines organise the enormous amount of information on the internet into simple answers to clear cut, search term-based, questions. What is deceptive about this everyday process is that it flattens rather than deepens and improves knowledge. Research questions which require more than immediate information are thwarted. For instance, we can easily find documents on the Web about Lucy, the hominid, but we have no direct way to discover the locations of finds similar in kind. Even though information on the web is densely linked - the average distance between documents is only 7 successive links [2] - the information itself is not related in a meaningful way. Hypertext links are made for human readers, rather than for machine interpretation. Digital libraries have not overcome the traditional paradigm of delivering a document as ultimate objective. Carl Lagoze states that "..the underlying public key infrastructure that was seen as 'essential to the emergence of digital libraries' remains undeveloped. Despite efforts of the W3C's Semantic Web initiative, the holy grail of semantic interoperability remains elusive" [8]

This paper argues that next-generation DL services must be built on accessing associations implicit or explicit in documents collection and their metadata. It suggests a new approach to leverage associations based on (i) generic core ontologies of relationship and co-reference links, (ii) semi-automatic maintenance of co-reference links by a new kind of service, and (iv) public engagement in the creation and development of the emerging network

2. ASSOCIATIONS AND IDENTITY

The ultimate goal of users is not to get an object but to *understand* a topic. Understanding is built on associations. Associations are found in digital objects or metadata. Metadata provide explicit associations in the form of relationships and data paths. Tools may extract associations from digital objects, either by interpretation of data structures or by statistical means such as evaluation of co-occurrence patterns, and save them again as metadata. Indices provide associations, and may also be seen as metadata.

The topic of associations has been faced both in the area of information retrieval and hypertext for many years and the following kinds of associations are widely used in Digital Libraries: Subject relations between documents and classes; subsumption of classes; hypertext links between documents; occurrence and co-occurrences of words. The latter two have weak semantics. There is a vast literature about statistical detection of associations in order to cluster documents by some co-occurrence patterns in the contents. They are mainly used to find similar documents, but *not to exploit the meaning* of the detected associations for understanding a topic. Ontology learning or automated thesaurus construction is a notable exception, but the semantics of the retrieved associations are generic (on a categorical level) and still very weak for subsequent reasoning. Even refined semantics of hypertext links have not brought any break-through in terms of topic-related automated reasoning so-far. It is hard to create powerful expressions from a combination of hypertext links for other purposes than getting documents and automatically following hypertext links readily retrieves the whole Web.

If the semantics of represented relationships are explicit, such as part-whole, membership, creation and participation, then patterns in the network of factual relations (or *material facts* [4]*)*, can reveal new, indirect associations, or can be used for inductive reasoning. There are many relevant applications, in which retrieval and discovery of digital objects themselves is based on simultaneous discovery of indirect associations, such as searching

First International Workshop on ``Digital Libraries Foundations" Vancouver, British Columbia, Canada, June 23, 2007.

for related literature based on co-citation [13], based on co-authorship networks ("friend of a friend", [5]), or search for business relations of dependent enterprises. Recently, Amit Sheth has stressed the extraordinary importance of access by factual relationships for the Semantic Web, in particular with respect to business applications [3]. The challenge is *not just to deliver* documents, but to *leverage on* the latent knowledge in the *combined* content of many digital sources.

Factual relations however can form meaningful semantic networks. In order to support any advanced services, relationships (i.e. classes of relations) should conform with a schema or ontology. Even though it is widely believed that there is no global ontology, the acceptance of Dublin Core demonstrates the opposite. If there is one or a few core ontologies, does not make any difference in their ability to give rise to global networks of knowledge. Empirical studies show [10] that the number of relationships in ontologies is orders of magnitudes smaller than that of classes and hence quite manageable. [6], [7], [14] have shown that a core ontology of ten to a hundred relationships can capture semantics of data structures across many domains.

Now, little advanced reasoning can take place if the elements of the network are not connected. They connect through the domain and range values of the relations that identify items in a domain of discourse. The identifiers are normally not unique and therefore don't match. This "duplicate removal" or *co-reference* detection [9] as co-reference is a process widely underestimated in importance for information integration. What actually relates propositions and other contents found in the documents *is not a "hyperlink"*, but the fact that they refer to the *very same items*. These may be events, dates, places, persons, material or immaterial things such as texts, images, names etc. Even terms can often be seen as (conceptual) items of discourse, rather than as expressions of classification. We argue that the actual semantics linking items are *in* the document, and not between them.

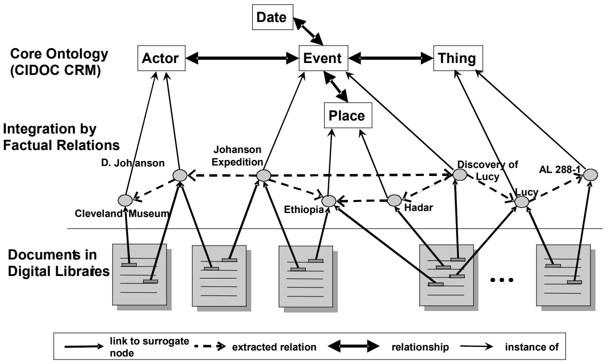
So the key to more advanced services seems to be the unique identification of things. The "bad news" is the immense number of things referred, orders of magnitude larger than the number of terms. We suggest a completely different approach: In order to connect facts, an automated system needs not know any detail about the referred items besides that they are identical. *Wherever* the knowledge comes from, it does the job. So, equivalence clusters of explicit co-reference links between respective document parts or elements of database records can replace maintenance of identification data as traditionally done in authority files. This approach is more general, since the former can be generated from latter, but not vice-versa. Therefore we propose a new kind of DL service: *Co-reference Services (CRS)*.

3. ABOUT CO-REFERENCE INFORMATION

Librarians and others have invested heavily in so-called authority files or knowledge organisation systems (KOS) [12], which register names and characteristics of authors and other items and associate them with a preferred representation in a central resource, and then advise colleagues to use the central resource as a reference to obtain unique identifiers. But using a central resource causes serious scalability problems. The standardization process always lags behind reality. Computer scientists tend to regard the recognition of co-reference (duplicate detection) as a question of probability that two items are referred to by similar names or similar properties (e.g., [1]). What is common to both approaches is the fact, that they do not preserve actual knowledge that an identifier a1 in source s1, and an identifier a2 in source s2, refer to the same real-world item. Only very recently, the project VIAF [11] has engaged in correlating two authority files with some nine million person descriptions into what they call a "virtual authority file" by a kind of co-reference links.

If we make the assumption that the maintainer or creator of s1 knows what a1 means, and the maintainer or creator of s2 knows what a2 means, both could *convene* and record the fact of co-reference without any common attribute or authority file. Philosophically, there is only one primary source for the identity of something: a citation in a document or data record field, i.e. "what the author meant by this expression". An record in an authority file poses the same question. All other questions of identity can be seen as elements of the subsequent co-reference problem. If the authors cannot be queried, one may base assumptions about co-reference on known common features of the citations under investigation. Those features may be based on values, such as a common name for the birthplace of a person, which are in turn subject to a co-reference question. Automated data cleaning methods work on the latter base.

Figure 1. Insert caption to place caption below figure



Obviously, co-reference is a question of belief based on explicit or implicit knowledge and evidence. Therefore we regard a co-reference statement as an elementary piece of scientific or scholarly knowledge, regardless of any heuristic-based software assisting in the identification process. Each co-reference statement allows for the connection of all factual relations to the two identifiers involved.

Intuitively, co-reference should be transitive and form equivalence classes that could scale up to any size. In order to relate the elements of an equivalence class of cardinality *v*, a minimal number of *(v-1)* primary equivalences is needed to derive all *v(v-1)/2* equivalences. This demonstrates the economic power of preserving co-reference knowledge once the networks grow tighter. Each equivalence class can be regarded as a *digital surrogate node* for the referred item. The global number of *surrogate nodes* per real item may be used as an inverse measure for the degree of integration of knowledge sources.

So, if we *publish* a co-reference statement and preserve the referential integrity, we have achieved more than any authority file: we have connected facts from two information assets *to our best knowledge.* (See figure). In contrast to hypertext links, this information can have a tremendous impact on computer-supported reasoning. A major short-coming of query mediator approaches [15] to information integration is the difficulty to match identifiers on-the-fly. Data warehouse approaches or metadata harvesters are more flexible in this respect, but not as scalable. Explicit co-reference information could close the gap and allow for highly performant hybrid information integration system, i.e. configurations seamlessly including physical and virtual integration systems of metadata

4. CO-REFERENCE SERVICES

We have started to elaborate theoretical foundations for co-reference services, which will be published soon. It has also been subject of several recent applications for European research grants. We present here the general requirements for the envisaged services:

1. A Co-reference Service should be based on common protocols and standards for information access and integration. Webservices in a data GRID environment could provide a beneficial environment.

2. Co-reference links should be persistent and public so that investment pays off. They may be bidirectional or unidirectional. In the latter case harvesting should be foreseen to create the appropriate inverted indices (see 7.). The use of preferred identifiers from an authority file or gazetteer can be seen as a special case of unidirectional linkage, as long as their persistency is guaranteed.

3. Primary Co-reference links should be provided and maintained (curated) by teams having the expertise to assess their correctness, such as librarians, archivists, scholars scientists. Therefore they should be preserved in local, distributed databases (“indices”).

4. Social tagging should mobilize the potential of general users and domain experts to enhance and verify co-reference information. Scholars use to spend a large part of their research efforts to collecting and verifying co-reference information. Not all co-reference information is relevant. Social tagging can also create an emergent notion of relevance.

5. Co-reference links must be associated with belief values. Experts distinguish belief values, and trust in sources may differ. Belief values should be used to control precision and recall of retrieval following co-reference links.

6. Duplicate-detection algorithms can be used to populate co-reference indices. Appropriate belief values should distinguish automated from manual sources. Generic Webservice protocols and formats could be beneficial to run intelligent duplicate detection in GRIDs. Duplicate detection algorithms can benefit from co-reference indices.

7. The envisaged open environment requires global coordination: providers may publish bad information, they may not agree, information may be abandoned or relevant areas not covered. Global supervision can be done by open consortia setting the rules and doing central services for appropriate communities. They constitute the co-reference service in the narrower sense. The consortia should in turn collaborate on common standards. Central services are in particular:

    a. Controlling referential integrity and negotiating solutions with primary information providers. Maintaining inverted indices.

    b. Determination of the transitive closures of equivalence clusters. Detection of contradictory information and identification of possible sources of inconsistency. Duplicate detection algorithms can be modified to validate manual co-reference information.

    c. Guiding and monitoring work of primary information providers to conflict resolution, handling of abandoned sources, suggestions for new areas to cover. The employment of authority files can simplify complex co-reference clusters. The service can feed into authority file maintenance.

5. CONCLUSIONS AND FUTURE WORK

We have argued that a next generation of DL systems should leverage on associations across document contents, metadata, indices and collections. We regard explicit co-reference information as enabling factor of great genericity and propose a new kind of DL service integrating data cleaning methods and reference information management in KOS. It has the potential to open up radically new applications on top of DLs. Reasoning services long dreamed of may become feasible in the envisaged connected knowledge networks. To our opinion, the whole area deserves a major research effort. DL research focus should shift from classification to association. We continue research on foundational issues and algorithms for consistency verification and maintenance of co-reference information: How can global consistency be improved in a distributed system? What are the integrating and disintegrating factors?

6. REFERENCES

[1] Bilenko, M., and Mooney, R.J. Adaptive Duplicate Detection Using Learnable String Similarity Measures. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, (Washington DC, August 2003), 39-48.

[2] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. Graph structure in the web. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, *33, 1-6*, (2000) 309–320, ISSN: 1389-1286.

[3] Cardoso, J., and Sheth, A. Eds. *Semantic Web Services, Processes and Applications*. Springer, 2006, 405 pages, ISBN 0-38730239-5.

[4] Degen, W., Heller, B., Herre, H., and Smith, B. GOL - Towards an Axiomatized Upper-Level Ontology. *Electronics and Computer Science* (2001).

[5] Dodds, L. *An Introduction to FOAF*. 2004. At http://www.xml.com/pub/a/2004/02/04/foaf.html, accessed Nov.16, 2006.

[6] Doerr, M. The CIDOC CRM - An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine, 24(3*) (2003).

[7] Doerr M., Hunter, J., and Lagoze C. Towards a Core Ontology for Information Integration. *Journal of Digital Information, 4, 1 (*2003) Article No. 169.

[8] Lagoze, C., Krafft, D. B., Payette, S., and Jesurogai, S. What Is a Digital Library Anymore, Anyway? *D-Lib Magazine*, *11, 11*, (2005).

[9] Levesque, H.J. Foundations of a functional approach to knowledge representation. *Artificial Intelligence, 23, 2,* (1984), 155–212.

[10] Magkanaraki, A., Alexaki, S., Christophides, V., and Plexousakis, D. 2002. Benchmarking RDF schemata for the Semantic Web. The Semantic Web - ISWC 2002: In *Proceedings of the First International Semantic Web Conference* (Sardinia, Italy, June 9-12, 2002). Springer Berlin / Heidelberg 2342/2002, ISSN:0302-9743.

[11] O'Neill, E.T., Bennett, R., Hengel-Dittrich, C., and. Tillett, B., B. *Viaf (Virtual International Authority File)*: Linking Die Deutsche Bibliothek And Li-Brary Of Congress Name Authority Files. In WLIC2006, 2006.

[12] Patel, M., Koch, T., Doerr, M., Tsinaraki, C., Gioldasis, N., Golub, K., and Tudhope, D. *Semantic Interoperability in Digital Library Systems*, DELOS Network of Excellence on Digital Libraries – deliverable 5.3.1, June 2005.

[13] Salton, G. Associative Document Retrieval Techniques Using Bibliographic Information, In *Journal of the ACM*, 10 (1963), pp 440-457

[14] Sinclair, P., Addis, M., Choi, F., Doerr, M., Lewis, P., and Martinez, K. The use of CRM Core in Multimedia Annotation. In *Proceedings of the 1st First International Workshop on Semantic Web Annotations for Multimedia part of the 15th World Wide Web Conference (SWAMM 2006)* (Edinburgh, Scotland, May 2006.)

[15] Wiederhold, G. Mediators in the Architecture of Future Information Systems. *IEEE Computer*, (March 1992).