International Journal of Digital Content Technology and its Applications
Volume 4, Number 3, June 2010

Practicability of Dataspace Systems

Hamid Turab Mirza, Ling Chen and Gencai Chen*
College of Computer Science, Zhejiang University, Hangzhou, 310027, China
hamid306@zju.edu.cn, lingchen@zju.edu.cn, chengc@zju.edu.cn
doi: 10.4156/jdcta.vol4.issue3.23

Abstract

Nowadays there is rarely a scenario where all the data can be fit nicely into one relational database management system or any other single data model. In acknowledgement of this fact a new concept of Dataspaces was introduced, according to which a dataspace system is assumed to be a hybrid of a search engine, a databases management system, an information integration system and a data sharing system. However the concept was presented in a visionary way and its implementation in the real world scenario has opened up many complex research challenges. Moreover so far the efforts put forth by the research community are quite disparate and it is highly desirable to have a unified effort which would hopefully enable rapid progress. Furthermore due to very high end-user expectations of such systems there are a lot of challenges and problems that need to be resolved and much scope for future work remains.

Keywords: *Dataspace, Information Integration, Unstructured Data, Database, Indexing*

1. Introduction

The broad field of data management, which is generally inferred to include all types of data, has grown tremendously in the last ten years or so. Nowadays anybody with an access to a computer deals with a vast variety of data, Doan et al. [1] points out that especially in large organizations more and more people are dealing with disparate forms of data to accomplish their everyday tasks. Specifically the most prominent kind of data is unstructured, which encompasses text documents, emails, web pages, memos, call center text record and alike. Halevy et al. [2] state that, according to an estimate, this kind of data is currently 80% of the world data, and is growing rapidly. Which if managed in a principled way, can offer substantial benefits to everyday users and the organizations.

Many other researchers [3, 2, 4] have also pointed out that in current data management scenarios, it is very rare to have a situation where all the data in an organization can be managed within a single database management system or any other independent data model. This is because most of the data is not very well structured and hence remains outside database systems. And now it is quite obvious that the key database players are not being able to play much role in managing it. Moreover since unstructured data is growing at a rapid pace, naturally to perform well; more and more software systems need access to both structured and unstructured data.

To cater for the above mentioned setting; in December 2005, Halevy et al. introduced a concept of Dataspaces. And acknowledged the fact that the data management challenges we face are at a much higher level and we need to manage a dataspace, rather than a database. Franklin et al. [5] and Halevy et al. [2] further elaborate that a dataspace consists of a set of participants and a set of relationships. Participants are individual data sources and could be relational databases, XML repositories, text databases, Web services, data-stream systems, sensor deployments, or any other element that stores or delivers data. A dataspace should be able to model any kind of relationship between two (or more) participants. In the extreme case, a relationship is a full schema mapping that enables arbitrary data exchange and query reformulation among participants. In other cases, the relationship can express simple dependencies, where the details are not known precisely. Examples of dataspaces include: an enterprise, the desktop, a library, large scientific projects or a smart home.

* Corresponding author.

However Elsayed et al. [6] argue that the dataspace concept is presented in a visionary way, and its realization in actual application environments has opened up new research challenges. These research challenges can be broadly put into the following two sets. The first set of challenges includes finding related data sources, provision of search and query facility, tracing lineage and determining correctness of data. The second set is primarily concerned with the administration side and includes challenges of rule enforcement, applying integrity constraints, ensuring that naming conventions are followed over the whole collection, making sure that data is available, devising mechanism for data recovery, access control and managing the changes in data and associated metadata.

Sheer complexity of the above mentioned set of challenges has raised some concerns about the viability of the dataspace systems to be. In the rest of the paper; first of all practical importance of the dataspace systems with respect to the industry needs is highlighted; which is followed by the discussion of the user expectations of such systems. Latter on major research challenges faced while trying to meet the user expectations are pointed out; and in the end different prototype systems are evaluated and discussed.

2. Dataspace Systems and Industry

We have experienced in the past that although working on the technical aspects of the concept is all well and good, however working on technical aspects alone is not enough for success. Moreover it is stressed [1] that for any theoretical idea to materialize it should have some associated business target. One of the reasons for this argument is financial, as the industry can support the research work, however even for non-financial reasons there is a need of a business target so that research, development and production cycle can be created among the industry and the academia. Moreover it can be seen that the presence of a successful database management industry has played a pivotal role in the success of database research. Therefore it is anticipated that an equivalent industry will be essential for going forward in the area of dataspaces too.

However, providentially the concept of dataspaces is inspired in large part by the challenges faced by industry today. In fact, there are many examples where industry is already making steps in this direction, but Franklin et al. [5] have raised the concern that these steps are isolated from each other and there is a clear need for a broader view that will yield a cleaner system abstraction and set of techniques. For example Chu et al. [7] have proposed that ideally while interacting with a dataspace system at any point in time, a user should be able to query using as much or little structure as is currently known, and should be able to perform increasingly sophisticated queries as the system incrementally evolves its understanding of the data. However unfortunately, up till now there is little consensus about what kind of system should be used to support this kind of incremental capability [7, 5].

In particular it is argued that to effectively manage a hybrid mix of data, a clear model of how data is generated and exploited should be developed and a complete system that implements this model should be designed accordingly. It is hoped that this type of system will help the research community to make a combined effort. Finally, it is argued that there is a strong need of an accompanying business community that can ensure a cycle of "ideas to realistic prototype to commercial transfer back to ideas" [1]. Nevertheless there are huge user expectations from such systems and delivering promising prototypes will be an enormous challenge, however; it is anticipated that without the support of associated industry the research community might not reach its potential.

3. User Expectations

As touched upon previously, in today's data environment users expect a refined response when queries similar to the following are posed over a collection of structured, semi-structured and unstructured data:

- Find me the business proposals discussed at last year's ABC conference and the presentation slides (maybe in an attachment).
- List me all the documents sent to associated companies regarding the same conference.
- List me the phone numbers of my business affiliates with whom I had meetings in the last 2 months.

- Find me the contact addresses of people, who have sent me any emails in the last 15 months.

However up till now it is not possible for the web search engines, database systems, desktop keyword search programs or any other isolated application to answer all the above queries in a graceful manner. Whereas the very concept of a dataspace system [2] is that it should provide base functionality over all data sources and respond reasonably well to the queries that are comparable to the queries mentioned above. For example, a dataspace support platform (DSSP) should provide keyword search over all of its data sources, similar to that provided by existing desktop search systems. However when more sophisticated operations are required, such as relational style queries, data mining, or monitoring over certain sources, then additional effort should be applied to more closely integrate those sources in an incremental, "pay-as-you-go" fashion [2].

Notably one of the main reasons of the success of web search systems is that users do not expect the results to be accurate; furthermore if they get something even close to what they are searching for then they are not much bothered about all the other irrelevant links and information presented to them. Here Jagadish et al. [3] raises a question that, whether it will be acceptable for the user to have a search engine kind of experience with the dataspace systems, answer to which is obviously 'NO'. This is because of the different characteristics and properties of dataspace system that distinguish it from web search engines, database systems and desktop search applications. Consequently users have fundamentally different expectation from the dataspace systems.

One of the main expectations from dataspace systems is that users want to query the system in a refined way. Although when using the web search users are happy with keyword search, however the same users when interacting with dataspaces expect accurate data in response to their queries. This is simply because of the fact that they know that underneath there is rich relevant data residing in a much more organized way. This kind of response cannot be generated from simple keyword search. These high expectations raise many issues; a response quite similar to the database system is expected and furthermore for any unexpected result the system should be able to explain the user that 'where did this result came from?'. Jagadish et al. [3] have stressed that to cater for above mentioned expectation, provenance tracking will be a necessary feature in dataspace systems.

Moreover while doing a web search a user wants to see distinct links to the websites and does not expect those links to have any relationship among them. On the contrary, in the case of dataspaces, in response to a particular query a user might expect to see a picture, an e-mail associated with that picture, the attachment in that e-mail and alike. Moreover to worsen the scenario, user's expectation depend quite largely on the domain in which he is working and users mental model of the system. Another issue raised by [3] is that the web search engines are only meant to do searching whereas dataspace systems by their very nature will need to be updated with time, as users will want to store new or remove old data from their dataspace, therefore it must cater to these needs also.

In pursuit of meeting the above mentioned user expectations and in an effort to make dataspace systems absolutely practical; the research community is working on various research agendas and problems. Key challenges and their proposed solutions are discussed in the following section.

4. Key Research Challenges

Halevy et al. [2] have identified that, the following features differentiate dataspace systems from conventional database management and data integration systems:

- A dataspace system should be capable of handling data and applications in different formats accessible through many systems with different interfaces. Unlike database systems, it cannot leave out some of the data and must deal with all kinds of data in a dataspace.
- Although a dataspace system is suppose to provide full functionality of searching, querying, modifying and deleting the data. However the same data might be accessible by many other systems and application, which can update or modify the data on their own. Therefore unlike database systems dataspace systems do not have a complete control over their data.
- When a query is posed to a dataspace system, it is suppose to offer varying levels of services and may only be able to respond with a best-effort or an approximate answer depending on the availability of the data sources at that time.

- If and when required a dataspace system should be able to create tighter integration of data in an incremental fashion.

due to the above mentioned distinguishing properties of dataspace systems, currently research community is faced with the following key challenges of; 'integration of heterogeneous schemas', 'indexing of loosely coupled collections of data', 'seamless querying on structured and unstructured data' and 'soliciting user feedback'.

4.1. Integration of Heterogeneous Schemas

The goal of data integration systems is to offer a uniform interface to a set of data sources. Recent research [8, 9] has proposed new architectures for dealing with large scale heterogeneous data integration. However despite recent progress, setting up and maintaining a data integration application still requires significant upfront effort of creating a mediated schema and semantic mappings from the data sources to the mediated schema. To completely automate data integration, we need to automatically create a mediated schema from the sources and automatically create semantic mappings between the sources and the mediated schema.

Generally dataspace systems are viewed as the next step in enhancement of data integration systems. However Halevy et al. [2] have argued that dataspace systems are different from current data integration systems. They further elaborate their point that although data integration applications interact with disparate data sources that may exist on different systems, however before providing any services to the user these systems must know the precise relationship among all the schemas and sources. Hence setting up the data integration system requires labor intensive upfront effort. This is the main distinction between data integration and a dataspace system, that in DSSPs no upfront effort is required and system provides the basic functionality from the very start.

According to Sarma et al. [9] the ultimate goal of dataspace system is to do data integration without any human involvement. The resulting integration should give best-effort answers, and should let the administrator improve the system in a pay-as-you-go fashion. Though it would be really good to achieve this goal however; Dong and Halevy [10] doubt these capabilities and state that, although a dataspace system typically employs automatic methods that try to extract some of the semantic relation-ships, but these results are approximate at best. Furthermore to make things more complex in some cases, semantic relationships are unknown because of the sheer number of sources involved or the lack of people skilled in specifying such relationships. And in other cases, not all semantic relationships are necessary in order to offer the services of interest to users.

In data integration applications, the two main activities that require significant human effort are creating the mediated schema and semantic mappings between the data sources and the mediated schema. This is because of the fact that these two activities depend on the domain knowledge and understanding of the queries that can be asked frequently. Moreover in many scenarios there is no need to integrate all the sources for the system to be useful. This motivates the pay-as-you-go approach to integration, which means that in the beginning the system can provide services with very few semantic mappings; which as and when required can be improved in an incremental fashion [9]. These features of dataspace systems make the idea of absolute automatic data integration a challenging task.

Efficiency and data quality are two main standards for evaluating integration strategies, previous and current works under the dataspace agenda have been trying to achieve both. PAYGO by Madhavan et al. [11] is a system for integrating structured and unstructured data on the web. In the same way [12] in their work on CIMPLE project have developed a system for data integration from online communities. As an extension to these works, ROOMBA [8] assists these systems in getting the feedback from users. Notably Sarma et al. [9] have demonstrated techniques to automatically create a mediated schema and semantic mappings to the mediated schema. Different from other works, the features of user access behaviors are paid special consideration to in OrientSpace system. Moreover Li and Meng [13] suggest that by utilizing the pattern of user access, data integration will become more automatic and efficient.

Despite all the efforts there are still many open challenges and data integration for dataspace systems is still a work-in-progress, for example it is still an issue that how to effectively mine the

pattern as suggested in the OrientSpace system. Moreover very recently Dong et al. [14] have introduced the notion of probabilistic schema mappings which provides a foundation for answering queries in a data integration system with un-certainty about semi-automatically created mappings. Moreover concept of Object Weight is introduced recently, Li and Meng [13] point out that by experiments the reasonability of the Object Weight and that it is a main parameter effecting operation efficiency is proved. Nevertheless how to formulate and compute it is still a problem, as it depends on many factors, such as data type, created time, modification time, data size, and so on.

4.2. Indexing of Loosely Coupled Collections of Data

Index is an efficient way to improve query performance. In context of dataspace systems the ultimate goal is to support robust indexing of loosely coupled collections of data in the presence of varying degrees of heterogeneity in schema and data, such that the system can efficiently answer queries that combine keywords and simple structural requirements [13, 10].

Ideally a user should be able to search and query a dataspace through an interface. Dong and Halevy [10] suggest that while designing such a system, the following should be considered. First, the user's interaction with a dataspace is of exploratory nature and secondly there is no single schema to which user can put queries to. Therefore it is important that the user should be able to specify both structured and keyword queries. Moreover the results should contain not only the matched data but should also present the objects associated with that data.

The existing methods for dataspace systems have either opted to build a separate index for each attribute in each data source to support structured queries on structured data, or have created an inverted list to offer keyword search capabilities on a hybrid mix of data. Li and Meng [13] make a point here that creating a full text index will result in a large size and will eventually create efficiency problems. Hence it is argued that this method will fail to support queries that combine keywords and structure. In context of XML much work has been done on indexing structures and keywords, however the following two reasons suggest that the techniques for XML are not appropriate for dataspaces. First, techniques for XML primarily rely on modeling the parent-child and ancestor-descendant relationship; which is not always the case for dataspaces. Secondly XML indexing techniques rely on building multiple indexes; and visiting multiple indexes to answer a predicate query or a neighborhood keyword query can be quite time consuming.

Dong and Halevy [10] have described an indexing technique to support flexible querying in dataspace environment. Their technique allows the users to pose structured queries when they can and also allows them to fall back to keyword search when they are unable to exploit the underlying structure. Furthermore their technique extracts not only the results where the keyword matches but also the objects associated with it. Their method extends inverted lists to capture structure when it is present, including attributes of instances, relationships between instances, synonyms on schema elements, and hierarchies of schema elements. While experimental results by Dong and Halevy [10] have shown that incorporating structure into inverted list can improve the efficiency of query answering; however it is argued that still there is a need to extend the index to support value heterogeneity and more enhancement can be achieved by investigating appropriate ranking algorithms in context of DSSPs.

Another notable work is by Li and Meng [13], in which they have proposed a two-level index strategy, the first level is index of CoreSpace, and the second level is index of the full space. Different policies are employed in maintaining the two kinds of indexes. However the effectiveness of this idea needs to be tested, similarly currently several methods for extending inverted lists are being worked at by the research community and still there is a need of experiments to validate the utility of these techniques.

4.3. Seamless Querying on Structured and Unstructured Data

One of the key services of a DSSP is to provide seamless querying on the structured and unstructured data. Moreover much of the user interaction with dataspaces involves exploring the data, and users do not have a single schema to which they can pose queries. Consequently, it is important

that queries are allowed to specify varying degrees of structure, spanning keyword queries to more structure aware queries [4, 15, 10].

Until now querying each kind of data in isolation has been the main subject of study for the fields of databases and information retrieval. Recently the database community has studied the problem of answering keyword queries on structured data such as relational data or XML data. Liu et al. [4] record that the only combination that has not been fully explored is answering structured queries on unstructured data. In their work they have taken the first step towards extracting keywords from structured queries even without domain knowledge and have proposed several directions which can be explored to improve keyword extraction when domain knowledge exists. The ability to widen queries in this way is an important capability in querying dataspaces, which includes heterogeneous collections of structured and unstructured data.

Similarly to improve the efficiency of query in dataspace systems, Li and Meng [13] have proposed the following technique. Based on their Query Framework, the user can efficiently perform a query by scanning the CoreSpace, instead of scanning the whole dataspace. And only when the user is not satisfied with the result, the full space is scanned. Furthermore their system is able to improve query efficiency, but there are still many challenging problems that need to be resolved in their proposed framework, such as how to define the boundary of the CoreSpace, how to define and formulate the satisfaction of user, and so on.

Although many experimental results have already shown that the above mentioned techniques have obtained good results in various domains, however still there are multiple directions for future work [4]. First, the extracted keyword sets could be refined by considering the schema or maybe even a corpus of schemas. For example, an extracted keyword can be replaced with a more domain-specific keyword in the schema; furthermore keywords selected from the corpus can be added to further narrow down the search space. Moreover existing structured data can be used, as proposed in SCORE [16], to supplement the selected keyword set. Also linguistic analysis of the words can be performed in the structured query to determine whether they are likely to be useful in keyword queries. Finally, it could be very useful to develop methods for ranking answers that are obtained from structured and unstructured data sources.

4.4. Soliciting User Feedback

As discussed above one of the main challenges of dataspace systems is of data integration. Normally a dataspace system will employ different techniques to automatically create schema matching [17] and entity resolution [18] between disparate data sources. And expected output of this is a set of potential matches, confirmation of which can dramatically improve the performance of a dataspace. However [8, 19] have stressed that it is very necessary to introduce human involvement at this stage and feedback from user could be of much benefit.

Doan et al. [1] also argued that if we are to be successful, our data management model should be designed to allow human intervention at key points of the end-to-end data management process. However one of the challenges is that in most of the scenarios there are a lot of possible matches that need to be confirmed by the user; however the user is the scarcest resource and certainly cannot be asked to confirm all of them. Jeffery et al. [8] identified that, here comes in the main principle of dataspace systems that is pay-as-you go, which initially allows the system to run on the basic functionality and then over the time with the feed-back of the user improves the whole process in an incremental fashion. Another primary challenge in this regard is that, when asking the user for confirmations the system should present the candidate matches in an order that is of maxi-mum benefit to the dataspace system [20, 21].

So far the work on getting the user feedback has focused on single mechanisms, especially in data integration scenarios the research carried out by [22] and [23] has tackled the problem by assisting the schema matching task with user feedback. Similarly, [24] have worked on an active learning based approach in which the feedback from the user is fed into the system to train the classifiers. However all these approaches are primarily based on only one type of data integration task and their method of selecting the potential matches for presenting to the user are heavily dependent on the type of classifiers used. Furthermore in case of dataspace systems the main drawback of these techniques is that; there main concern is to reduce the uncertainty of the matches where as there is a need to evaluate

these matches on the basis of the benefit to the overall system. Therefore it is argued that instead of different mechanisms that might work very well in their own capacities, there is a need of a system that combines them all and then makes the decisions which could possibly serve the goal of providing improved query results for the dataspace.

Notably the data integration technique taken up by MOBS [20] is to confirm the candidate matches by collecting feedback from many users and then making a decision on the basis of the combined result. Though MOBS is one step ahead from other systems, since it combines multiple mechanisms; however it lacks in the area of putting up potential maximum benefit matches to the users for their confirmation. On the other hand, the approach taken up by Jeffery et al. [8] works very well with the MOBS system. Both systems communicate with each other and ROMBA assists the MOBS system in deciding which candidate matches should be presented to the user. Despite the fact that we already have a few examples where reusing human attention has been very successful, however Halevy et al. [25] have stressed that this is an area that still has a lot of scope for additional research and development.

In the following section different dataspace prototype systems have been appraised and their lacking of the desired features, if any, is discussed.

5. Prototype Systems

The relational world received a huge benefit from the early creation of complete prototype systems such as System R and Ingres. With these systems as examples and context, an entire community arose working on improving their performance and broadening their scope. Doan et al. [1] make a point that, this unified a lot of what would otherwise be disparate work, helped guide research, enabled rapid progress, and resulted in real-world systems that magnified the dissemination of the products of data management research community efforts. Today in the complex environment of structured, semi structured and unstructured data, it is proposed that it is highly desirable to have similar example systems, those which can rally the community and unify the work, and hopefully enable rapid progress.

In context of dataspace systems many isolated efforts have already been made, one of the major works has been done within iMeMex system [26, 27]. iMeMex is developed with an aim to be a unified solution to personal information management and integration. It is designed to integrate seamlessly into existing operating systems like Windows, Linux and Mac OS X. Moreover it enables existing applications to gradually dispose file based storage. By using iMeMex, modern operating systems should be able to make use of sophisticated DBMS, IR and data integration technologies.

Another notable project is SEMEX [28] which on the basis of meaningful objects and associations creates a logical view of user's personal information. The users of this system are able to explore their personal information by objects such as Person, Publication and Message and associations such as AuthoredBy, Cites and AttachedTo. Keeping in view the fact that users normally do not want to take any pain to create any extra structure in their personal information, the Semex system attempts to create the association database automatically.

OrientSpace proposed by Li and Meng [13] is another prototype system developed for personal data integration and management. Based on CoreSpace framework and vertical data model, OrientSpace implements two functions: data integration and data query. It aims to initially integrate most data items in personal desktop computer. Different from other systems, it aims to provide a comprehensive solution for personal dataspace management. It not only supports desktop query, but also support data integration, data update, data backup, and so on.

Other than these major consolidated efforts a number of isolated research projects have been carried out like; The LifeStreams Project [29] organizes documents in chronological order and allows the user to view the documents from different viewpoints in terms of time. The Placeless Documents [30] annotates documents with property/value pairs, and group's documents into overlapping collections according to the property value. The Haystack [31] and MyLifeBits [32] projects view personal data as a graph of information. Nodes in the graph represent documents and annotation metadata; edges represent the annotates relationship. Stuff I've Seen [33] emphasizes access through text search which is independent of the application storing the data. It indexes all types of information and provides a unique full text keyword search interface.

6. Scope for Future Work

The challenge of implementing dataspace systems represents perhaps the largest data management opportunity for data management research community since managing relational data [1]. Despite the fact that a lot of dataspace prototype systems are being developed or are considered to be a work-in-progress, however none of them can be considered as a polished dataspace system that can be up to the expectations of end users. In the following, limitations of the current systems are discussed and the areas where there is a room for improvement are highlighted.

As discussed above, in context of dataspaces researchers have already established a fairly advanced starting point for pay-as-you-go data integration systems. At its core, these systems are built on modeling uncertainty in data integration scenarios. However; Sarma et al. [9] argue that setting up the data integration is just the first step in the process. Aside from improvements to this process, there is a need to consider how to improve the data integration system with time. Moreover it is believed that the foundation of modeling uncertainty will help pinpoint where human feedback can be most effective in improving the semantic integration in the system. In addition, there is a need to extend the techniques to deal with multiple table sources, including mapping multi-table schemas, normalizing mediated schemas, and so on.

Other than that Halevy et al. [25] have pointed out that due to the challenging nature of data integration problems, it is quite sure that these issues will keep the research community busy for quite some time. They further state that since data integration strategies require people to collaborate and share data, therefore one of the major issues is not technical but social. Moreover it involves extracting the suitable data, convincing people to share it and to get users to do this; the system needs to show some incentives for doing so. Another issue is of ensuring the privacy and security of the shared data sources. To make things more complicated, in many real time scenarios it is not even clear what it means to integrate data or how combined sets of data can operate together.

Another direction where a lot of work needs to be done is; how to handle unsatisfactory user feedback? Up till now most of the works have considered that whenever a user is presented with the candidate matches he/she always responds correctly. On the contrary Jeffery et al. [8] stress that; this is quite a false assumption as human beings could not be cent percent all the time and the answer could be much less than perfect. Moreover there is also a possibility that the system posed the questions in a confusing way, and it confused the user. Therefore to cater for the uncertainty in user feedbacks, the dataspace system might consider asking the same question to many users and employing a majority voting scheme.

An interesting area for further work could be to explore different types of user feedback. A lot of work has been done on implicit feedback [34, 35, 36] and dataspace systems can get benefit from that. For example in query answering process correctness of matches can be determined with the clickthrough rates; a click on some specific result might help in increasing the system confidence that this result might be correct and as a result dataspace system can improve the ranking of those matches. Moreover information from succeeding queries or query chains [37] can also be used to assist the selection of the correct match.

In addition to the above mentioned directions, Franklin et al. [5] have pointed out that for further enhancement of dataspace systems it will be crucial to leverage techniques from several other fields also. For example, recent developments in the field of knowledge representation can offer following two main benefits for heterogeneous collections of data in a dataspace. First is the presence of simple but useful formalisms for representing ontologies, and second is the concept of uniform resource identifiers as a mechanism for referring to global constants on which there exists some agreement among multiple data providers. Moreover, because of the very nature of dataspace systems there is a certain degree of uncertainty about the data, its lineage, correctness and completeness. To our advantage; due to the inherent uncertainty in the field of AI researchers have already developed several formalisms for modeling uncertainty which can be used in this context; but the challenge is to find models that are useful yet simple, understandable, and scalable.

As touched upon previously, most of the data in a dataspace will be a mix of unstructured and semi-structured data. Hence use of Information Retrieval methods could be very helpful in development of

DSSPs. Moreover [5] have pointed out that due to the complexities of dataspace systems, users might not exactly know what they are looking for or may not be able to properly interpret the output of the system, and therefore techniques from Information visualization can be incorporated to enhance the DSSPs.

From the above discussion it is quite evident that, much scope for future work remains and virtually every aspect of dataspace systems can be "drilled down" upon to discover and evaluate alternative approaches. Moreover despite the fact that the technical approach has its own merits, merely working on techniques to integrate disparate data sources, to extract structure from unstructured documents and allowing for human inter-action will not be enough for success. Retrospectively looking back on some key components in the success of relational systems may provide some insight as to what else is needed. And then we can use this insight to direct our efforts to unify and consequently develop viable dataspace systems that could prospectively be used at a commercial level.

7. Conclusion

Nowadays it is very rare to have a data management scenario where all the data in an organization can be managed within a single database management system or any other independent data model. This is because most of the data is not very well structured and hence remains outside database systems. To cater for this situation the concept of dataspaces was introduced, however the implementation of this concept in real application environment has opened up new research challenges, and has raised some concerns about the viability of the dataspace systems. Moreover it is believed that a theoretical concept cannot be materialized in a vacuum without any associated target business use of the idea. However, advantageously the concept of dataspaces is inspired in large part by the challenges faced by industry today.

Dataspace systems have promised a lot and it is expected that DSSPs should provide keyword search over all of its data sources, similar to that provided by existing desktop search systems. However when more sophisticated operations are required, such as relational style queries, data mining, or monitoring over certain sources, then additional effort should be applied in an incremental fashion. Due to many distinguishing properties of dataspace systems and in particular its differences from the database management systems and web search engines, currently research community is faced with the challenges like; integration of heterogeneous schemas, indexing of loosely coupled collections of data, seamless querying on structured and unstructured data and soliciting user feedback.

Despite the fact that a few dataspace prototype systems have already being developed or are considered to be a work-in-progress, however none of them can be considered as a polished dataspace system that can be up to the expectations of the end users. And it is quite evident that, still much scope for future work remains and virtually every aspect of dataspace system can be drilled down upon to discover and evaluate alternative approaches.

8. Acknowledgment

This work was funded by the Natural Science Foundation of China (No. 60703040), the Zhejiang Provincial Natural Science Foundation of China (No. Y107178), Science and Technology Department of Zhejiang Province (No. 2007C13019).

9. References

[1] A. Doan, J. F. Naughton, A. Baid, X. Chai, F. Chen, T. Chen, E. Chu, P. Derose, B. Gao, and C. Gokhale, "The case for a structured approach to managing unstructured data", In Proceedings of the Fourth Biennial Conference on Innovative Data Systems Research, http://www-db.cs.wisc.edu/cidr/cidr2009/Paper_110.pdf, 2009.

[2] A. Halevy, M. Franklin, and D. Maier, "Principles of dataspace systems", In Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 1-9, 2006.

[3] H. V. Jagadish, A. Chapman, A. Elkiss, M. Jayapandian, Y. Li, A. Nandi, and C. Yu, "Making database systems usable", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 13-24, 2007.
[4] J. Liu, X. Dong, and A. Halevy, "Answering structured queries on unstructured data", In Proceedings of the Ninth International Workshop on the Web and Databases, http://db.ucsd.edu/webdb2006/camera-ready/paginated/05-122.pdf, 2006.
[5] M. Franklin, A. Halevy, and D. Maier, "From databases to dataspaces: a new abstraction for information management", ACM SIGMOD Record, Vol. 34, pp. 27-33, 2005.
[6] I. Elsayed, P. Brezany, and A. M. Tjoa, "Towards realization of dataspaces", In Proceedings of the 17th International Conference on Database and Expert Systems Applications, pp. 266-272, 2006.
[7] E. Chu, A. Baid, T. Chen, A. H. Doan, and J. Naughton, "A relational approach to incrementally extracting and querying structure in unstructured data", In Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 1045-1056, 2007.
[8] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, "Pay-as-you-go user feedback for dataspace systems", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 847-860, 2008.
[9] A. D. Sarma, X. Dong, and A. Halevy, "Bootstrapping pay-as-you-go data integration systems", In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 861-874, 2008.
[10] X. Dong and A. Halevy, "Indexing dataspaces", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 43-54, 2007.
[11] J. Madhavan, S. Jeffery, S. Cohen, X. Dong, D. Ko, C. Yu, and A. Halevy, "Web-scale data integration: You can only afford to pay as you go", In Proceedings of the Third Biennial Conference on Innovative Data Systems Research, pp. 342-350, 2007.
[12] A. H. Doan, R. Ramakrishnan, F. Chen, P. Derose, Y. Lee, R. Mccann, M. Sayyadian, and W. Shen, "Community information management", IEEE Data Engineering Bulletin, Vol. 29, pp. 64-72, 2006.
[13] Y. Li and X. Meng, "Research on personal dataspace management", In Proceedings of the 2nd SIGMOD PhD workshop on Innovative database research, pp.7-12, 2008.
[14] X. L. Dong, A. Halevy, and C. Yu, "Data integration with uncertainty", The VLDB Journal, Vol. 18, pp. 469-500, 2009.
[15] B. Howe, D. Maier, N. Rayner, and J. Rucker, "Quarrying dataspaces: Schemaless profiling of unfamiliar information sources", In Proceedings of the 24th International Conference on Data Engineering, pp. 270-277, 2008.
[16] P. Roy, M. Mohania, B. Bamba, and S. Raman, "Towards automatic association of relevant unstructured content with structured query results", In Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 405-412, 2005.
[17] E. Rahm and P. A. Bernstein, "A survey of approaches to automatic schema matching", The VLDB Journal, Vol. 10, pp. 334-350, 2001.
[18] W. W. Cohen, P. R. Kumar, and S. E. Fienberg, "A comparison of string distance metrics for name matching tasks", In Proceedings of IJCAI Workshop on Information Integration on the Web, pp. 73-78, 2003.
[19] A. Doan and A. Y. Halevy, "Semantic integration research in the database community", AI magazine, Vol. 26, pp. 83-94, 2005.
[20] R. Mccann, A. H. Doan, A. Kramnik, and V. Varadarajan, "Building data integration systems via mass collaboration", In Proceedings of the International Workshop on Web and Databases, pp. 25-30, 2003.
[21] L. V. Ahn and L. Dabbish, "Labeling images with a computer game", In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 319-326, 2004.
[22] A. H. Doan, P. Domingos, and A. Y. Halevy, "Reconciling schemas of disparate data sources: A machine learning approach", ACM SIGMOD Record, Vol. 30, pp. 509-520, 2001.
[23] W. Wu, C. Yu, A. H. Doan, and W. Meng, "An interactive clustering based approach to integrating source query interfaces on the deep web", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 95-106, 2004.

[24] S. Sarawagi and A. Bhamidipaty, "Interactive deduplication using active learning", In Proceedings of the eighth international conference on Knowledge discovery and data mining, pp. 269-278, 2002.
[25] A. Halevy, A. Rajaraman, and J. Ordille, "Data integration: the teenage years", In Proceedings of the 32nd international conference on Very large data bases, pp. 9-16, 2006.
[26] L. Blunschi, J. P. Dittrich, O. R. Girard, S. K. Karakashian, and M. A. V. Salles, "A dataspace odyssey: The iMeMex personal dataspace management system", In Proceedings of the Third Biennial Conference on Innovative Data Systems Research, pp. 114-119, 2007.
[27] J. P. Dittrich, M. A. V. Salles, D. Kossmann, and L. Blunschi, "iMeMex: escapes from the personal information jungle", In Proceedings of the 31st International Conference on Very Large Data Bases, pp. 1306-1309, 2005.
[28] X. Dong and A. Halevy, "A platform for personal information management and integration", In Proceedings of the Second Biennial Conference on Innovative Data Systems Research, pp. 119-130, 2005.
[29] E. Freeman and D. Gelernter, "Lifestreams: A storage model for personal data", ACM SIGMOD Record, pp. 80-86, 1996.
[30] P. Dourish, "The appropriation of interactive technologies: Some lessons from placeless documents", Computer Supported Cooperative Work (CSCW), Vol. 12, pp. 465-490, 2003.
[31] D. R. Karger, K. Bakshi, D. Huynh, D. Quan, and V. Sinha, "Haystack: A customizable general purpose information management tool for end users of semistructured data", In Proceedings of the Second Biennial Conference on Innovative Data Systems Research, pp. 13-26, 2005.
[32] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong, "MyLifeBits: fulfilling the Memex vision", In Proceedings of the tenth ACM international conference on Multimedia, pp. 235-238, 2002.
[33] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, "Stuff I've seen: a system for personal information retrieval and reuse", In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 72-79, 2003.
[34] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback", In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 154-161, 2005.
[35] M. Claypool, P. Le, M. Wased, and D. Brown, "Implicit interest indicators", In Proceedings of the 2001 International Conference on Intelligent User Interfaces, pp. 33-40, 2001.
[36] H. Lee, B. Lee, K. Park, and R. Elmasri, "Fusion Techniques for Reliable Information: A Survey", JDCTA: International Journal of Digital Content Technology and its Applications, Vol. 4, No. 2, pp. 74-88, 2010
[37] F. Radlinski and T. Joachims, "Query chains: learning to rank from implicit feedback", In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 239-248, 2005.