Emerging Open Agoras of Data and Information*

Yannis Ioannidis
University of Athens, Department of Informatics and Telecommunications
yannis@di.uoa.gr, http://www.di.uoa.gr/∼yannis

Abstract

Open Agoras *of information are distributed environments of independent systems, where seeking for information is similar to real-life searching for material goods. Interaction with these systems may occur in several unconventional modalities, user behavior may be personalized and context-dependent, system reaction may be unpredictable, and the information produced as a result may also be personalized and context-dependent, negotiable and of uncertain origin or quality. This paper explores Open Agoras of information and identifies some technical challenges that they raise*

1. Introduction

In the past few decades, data and information management systems have changed dramatically the way businesses and organizations operate. Such systems offer precise and secure ways for such enterprises to organize their information assets as well as stylized languages and interfaces for the rightful users to search for accurate answers to particular information needs. Around these systems, comprehensive information environments are established, which are closely related to diverse business processes.

There is a growing need, however, for information management systems that support a rather different kind of environment, where consistency, accuracy, and absolute quality guarantees are sacrificed to some extent for certain other desirable characteristics, such as flexibility, adaptability, and diversity. Such an environment will move away from the closed world of business processes and resemble the open world of commercial exchanges. When searching for material goods in everyday life, people shop around, try to identify the best choices, make tradeoffs between various characteristics of what they want to find, negotiate with the merchants or owners of the goods, and sometimes proceed with purchases even if they are uncertain of the origin or quality of what they buy. Furthermore, the entire process may follow dramatically different paths for different people, affected by personal styles of shopping, different preferences on the kinds of goods sought and their characteristics, and even by the particular moods of the people at the time, their location, or their reasons for pursuing particular items.

Forthcoming information systems should support very similar styles of interaction. That is, there is a need to establish *Open Agoras* of information, i.e., distributed environments of independent information systems, where seeking for information will be similar to real-life searching for material goods. Interaction with these systems may occur in several unconventional modalities, user behavior may be personalized and context-dependent, system reaction may be unpredictable, and the information produced as a result may also be personalized and context-dependent, as well as negotiable and of uncertain origin or quality. This short paper explores such Open Agoras of information, identifies some technical challenges that they raise, and presents some thoughts on how to address them.

We have identified eight key issues that characterize interactions with an Open Agora and require attention by the research community: *Uncertainty* - There is uncertainty with respect to several aspects of a request in such an environment: identification of appropriate resources; metrics for matching unconventional objects; quality of the result. *Quality of Service (QoS)* - There are several quality indicators for query results or the query processing process itself that are beyond the traditional response time or work: completeness, freshness, trustworthiness, etc. At any point, users need to make tradeoffs among these parameters, which have become important due to the uncertainty that exists in the process. *Negotiation* - Achieving optimal values for the above QoS parameters or satisfactory balance points among them requires that users (or underlying query agents) negotiate with the information resources they deal with. *Personalization* - Based on their backgrounds, different users are interested in very different information even when they interact with the system in exactly the same way. *Socialization* - The social context in which users operate

*Partially supported by the Information Society Technologies (IST) Program of the European Commission under the DELOS Network of Excellence and the BRICKS, DILIGENT, and DRIVER Integrated Projects.

1-4244-0803-2/07/$20.00 ©2007 IEEE.

affects their perspective on some of their requests, making them interested in the responses that would match the preferences of their circle. *Collaboration* - When working on a common project, users concurrently pose individual series of requests in a group setting, observe each other's actions, and discuss and merge the results based on their interests. *Contextualization* - Based on the context users are operating in, they may be seeking very different answers from the very same query. *Multi-Modal Interaction* - In addition to queries, users interact with these environments via continuous information feeds, browsing of information sources, and group activities with their colleagues, possibly all these combined together.

In the rest of the paper, we briefly touch upon each of the above challenges/characteristics of Open Agoras and discuss the technical issues that need to be addressed in order to make them a reality. We use examples from a general scenario that captures the main elements of information search in these environments. Iris is a young researcher who is investigating the different styles of folk jewelry that have been worn across Europe through the times. In addition to the traditional scientific material of her field, she accesses repositories that provide information on holdings of many museums, government properties, and regional cultural organizations across the continent. She uses automatic feeds of history and tourism magazine articles on new exhibitions and collections, as well as auction catalogs and even popular fashion magazines. These arrive at her office as multimedia documents and are often annotated by her. She stores documents and other objects of high interest as well as her annotations in a personal information base that she maintains, which she also shares with Jason, a colleague in a different institution who is working on traditional dance forms.

2. Uncertainty

Usually, conventional data systems operate under conditions of absolute *certainty*. All parameters of a query, the environment in which it is being processed, and its results are assumed to be fixed and undisputable. With respect to the query itself, the objects in the database and in the query selection conditions are well typed, and the object-matching operators have well-defined and deterministic truth values. With respect to the processing environment, I/O throughput is assumed to be known to query optimizers and constant during query optimization and query execution. Finally, with respect to the query results, they are assumed to be 100% sound and complete, most up to date, and originating from well-known and trusted sources.

Nevertheless, *uncertainty* is everywhere in real life, even in the most primitive laws of physics. Information management could not be an exception and uncertainty is one of the key characteristic of Open Agoras. Consider Iris's query for comparing some jewelry image with pertinent information in the Open Agora. Independent of the language used to express this query, there is much uncertainty at the query semantics level. There is uncertainty about the feature-sets that should be used to match the query image to those in the information sources. Are the typical "visible features", e.g., color histogram or texture, enough, or does one need to employ more metadata on the actual image content? Given a specific set of features, there is uncertainty on the degree in which two images match. Typical metrics used for this case are not necessarily capturing the perception that a user has about a match. Furthermore, matching strategies for compound objects, which consist of multiple heterogeneous sub-objects, each with its own semantics and rules for matching, are far from obvious. For example, how does a web page of a fashion magazine match with an auction catalog, taking into account the images they contain, the corresponding text, and their different layout? Going even further, how can objects of different types be compared and matched, e.g., an image of a jewel matching an article that talks about traditional costumes in a particular geographic region? Operating with this level of uncertainty on the meaning of a query is clearly missing in current systems. Indexing and processing algorithms need to be devised or revised to be able to handle such cases.

After its interpretation, Iris's query must be processed with much uncertainty with respect to its cost as well as result size, which arises due to (internal) systems limitations and (external) user activity. Clearly, there is a limit on the accuracy of cost functions and data statistics used by query optimizers. Furthermore, user activity causes fluctuation on the parameters of these functions or the statistics themselves. Such uncertainty is partially overcome through dynamic or parametric query optimization or other techniques. Still, there are several new relevant issues arising in the context of Open Agoras that have not been addressed so far. These are associated with uncertainty at the user level [3], in direct relation to *risk*, which is rather difficult to model, as different attitudes towards risk make people behave very differently under uncertainty.

The results of Iris's queries are of uncertain quality as well. Given the modus operandi of an Open Agora, the information sources from which these results are coming and the overall quality of those sources are, in principle, unpredictable. Sources may have declined to participate because of overloading, unavailability, or black-listing of Iris's IP address. Responding sources may or may not be well-known and trusted, and may have been content-rich or content-poor in the area of the query. Hence, there is uncertainty on the correctness of the query results, their completeness, their freshness, and other features. New techniques are needed for choosing information sources and optimizing and processing queries so that the quality of their

results with respect to some of these features, as well as features of the processing itself (e.g., execution time) is maximized.

3. Quality of Service (QoS)

When Iris buys vegetables, she does not make a contract with the greengrocer. If the vegetables are not as fresh as promised, in time, her trust is reduced and shops for vegetables elsewhere. In other cases, usually of major purchases, contracts exist that spell out everyone's responsibilities. When Iris is looking for information in an Open Agora, an implicit (or explicit) *contract* should be made with the information sources to describe the expected freshness, completeness, correctness, or time of delivery of the final result. Similarly to real life, any of the involved parties can unilaterally (or in mutual agreement) break the contract, in which case it has to compensate the other party for the damage caused by this action. Open Agoras should model QoS through the use of Service Level Agreement (SLA) contracts, which in real life, are different from "normal" contracts in the QoS premium paid, according to the risk/uncertainty of the requested service [2]. They also require new query processing and optimization techniques that take into account the different types of contracts and the corresponding contract-breaking consequences.

4. Negotiation

Similarly to real life, in Open Agoras, information providers and consumers have *asymmetric* knowledge on most parameters affecting their interaction. Iris may not know the motivation of some information sources (e.g., altruistic, competitive, or deceptive), the difficulties faced during information retrieval, or the quality of the result. In turn, these sources may not know Iris's preferences.

Given the above uncertainties in Open Agoras and the QoS tradeoffs that exist, queries and their results should be recognized as commodities and query optimization should be modeled as a trading negotiation process. Query answers and query operator execution jobs (or parts of them) should be traded in the network until deals are struck and contracts are "signed" with some information sources for specific levels of QoS. Such trading may also occur recursively, in the sense that some nodes may play the role of intermediaries between other nodes (subcontracting).

In comparison to trading negotiations for e-commerce [5], query processing and optimization faces unique new challenges that stem primarily from two reasons. First, queries have a complex structure and can be broken into smaller parts. This presents great opportunities for execution of query parts by different information sources followed up by the appropriate combination of their results. Finding the appropriate source in the Open Agora from which to obtain each piece of the relevant information corresponds to a query optimization problem that is beyond current technology. Second, the potential features of interest are many, with different algebraic and computational properties (e.g., there are ways to estimate the cost of an operator but quite hard to estimate the completeness of its result). Also, any subset of these features may be together the target of a multi-objective optimization process, combined with a host of policies in QoS contracts. Again, query optimization in such conditions remains to be investigated.

5. Personalization

When Iris and Jason meet at conferences and workshops, they often skip a session or two and walk around the city where the meeting is held, doing some shopping as well. Although they are very good friends, they have very different interests on items they consider important (books vs. music) and very different perceptions of quality and attractiveness for similar items (poetry vs. biographies, classic vs. jazz). Moreover, they have very different styles of approaching the shopping process itself (quick and goal-driver vs. relaxed and serendipitous), different levels of ability to negotiate with the merchant, etc.

Similarly, when "shopping" for information, Iris and Jason have very different interests, preferences, and styles. At times, they use different modes of interaction (direct queries vs. data browsing) and user interfaces, they formulate their queries differently, they have a different sense and interpretation of similarity for various kinds of objects, and different interests within unstructured or semi-structured objects, such as images and text. Furthermore, they trust different information sources and have different preferences regarding the query-time vs. result-quality tradeoff.

From the above, it should be evident that every aspect of information seeking in an Open Agora, including the issues raised in the previous subsections (i.e., dealing with uncertainty, achieving quality of service, and employing negotiation techniques) is subject to individual preferences and should be personalized [4]. Systems should maintain *user profiles* capturing the personality, background, interests, and other characteristics of their users. At all stages during their interaction with the Open Agora, different parts of their profile should be taken into account, affecting and modifying the process in several ways. Examples include completion of queries, selection of feature sets for similarity testing, choice of information sources, application of negotiation policies, and several others.

The technical challenges that arise in supporting such personalized information shopping are significant and quite diverse. Appropriate *user models* should be identified to capture the various elements of a human being's person-

ality that may impact the process. Although there is already extensive work on modeling some of these elements, e.g., preferences, there are several others that remain untouched, e.g., negotiation styles. In addition, for each relevant user model, *profiling* techniques need to be developed that will observe users during their normal interaction with the system, interpret their actions appropriately, and formulate their individual profiles accordingly by populating the model. Again, although mining user logs for profiling has been studied already quite extensively with respect to some specific user actions (e.g., click streams) and in the context of some specific user aspects (e.g., preferences of individual items, values, or terms), existing work represents just the tip of the iceberg. Furthermore, generating a single, cohesive profile from local ones collected for the same user at multiple information sources presents the usual difficulties of data integration as well as some specific ones that arise from the particular nature of profiles, e.g., dealing with inconsistent behavior at different sources with respect to likes and dislikes. Storage and indexing of profiles, as well as selection and retrieval of the appropriate profile parts in each case, are technical problems that require solutions also. Finally, putting the sophisticated user profiles that are expected to result from the above process into action requires rethinking current designs of user interfaces, query semantics interpretation, query processing and optimization techniques, information presentation and visualization, and other parts of the information-search cycle. As a concrete example, optimizing queries according to different risk profiles of individuals, establishing those profiles through observations, and taking into account the relevant QoS aspects in SLA agreements during query optimization and execution is an exciting research area that requires contributions not only from the database and, in general, the computer science community, but also from the psychology and cognitive science communities.

6. Socialization

As a good friend of Jason, Iris is interested in his opinion on specific items she buys and vice versa. Also, Iris is curious about items that Jason shows interest in and vice versa. Naturally, whether for social or scientific reasons, such opinion exchange is important for information shopping as well. If personalization implies using the user's own profile to customize a query or other interaction, socialization implies that other people's profiles should be used concurrently as well to affect the relevance of an information item. In addition, socialization involves using one's own profile on queries that others pose to learn from their interests. Clearly, the set of others' profiles and queries that someone has access to must be restricted based on access rights that have been granted according to users' privacy concerns. In addition, at any point, only the profiles of other users that have some affinity with the current user should be considered, where affinity may be defined through profile similarity or other association.

Socialization requires that all techniques used for various aspects of personalization should be extended and enhanced so that they may use multiple profiles, prioritize profiles according to their affinity to the current user, and fuse their use with normal processing. It also requires techniques for profile publishing that respects various privacy concerns as well as establishing profile similarity (or other association) through cross-user activity observations.

7. Collaboration

Shopping is often a group activity, with groups of friends (e.g., Iris and Jason) engaging in it together. Be it because of jealousy or admiration, such interactions usually result in better purchases (let alone the fact that they make shopping a more pleasant experience). In principle, information shopping can also be experienced in a group setting for exactly the same reasons as outlined above. Such collaboration is essentially socialization characterized by simultaneity as well. That is, it is realized through synergistic concurrent interactions of multiple (probably, a small number of) users with the Open Agora. They have a common goal but seek relevant information by exploring the market based on their individual profiles. They all see everyone's results at the same time, potentially fusing some of them into richer collections, and one may pick up on someone else's thread of actions and continue exploration based on one's own profile. Collaboration is essentially of the same spirit as *extreme programming*, only that it is not applied to program construction but to information exploration.

Such online, dynamic collaboration of users during information shopping is far from any conventional form of interaction with database or information-retrieval systems. In addition to the technical issues raised by socialization, collaboration also brings up several variations of the multiple query optimization problem where different user profiles are used for different queries.

8. Contextualization

The details of any human activity often depend on the general context within which they are performed. Context is a rather complicated concept with several dimensions, including time, location, general task performed, other people's presence, and immediately preceding activity [1]. For example, Iris buys wine in Paris, olive oil in Athens, and vinegar in Rome; she is very quick when buying clothes, always looking just for what she knows that she needs from

before, while she prefers to browse bookstores aimlessly in case she finds something interesting; she also watches thrillers but only when her children are not present.

Likewise, Iris approaches information shopping differently depending on the context. For example, she usually browses the Open Agora for information at the beginning of a research project but poses direct queries asking for specific items at the end of a project, when writing papers with her results; she also requires very different levels of similarity-operator accuracy and execution speed when looking for a place to eat versus when looking for a particular style of jewelry in recent auction catalogs.

Contextualization and personalization are mutually related in two distinct ways. First, the context of a person's activities may be seen as a generalization of the person's characteristics, i.e., a user profile may be seen as one dimension of the context. Second, the nature of personalization may be context dependent, i.e., someone's (active) profile may be different according to the context. More precisely, the profile of a person itself may include alternative choices for its various parts, with each choice activated when certain conditions hold.

Given the first relationship above, the technical challenges that arise in supporting contextualization in an Open Agora are similar to those of personalization. Appropriate *context models* should be identified to capture all relevant aspects of the context; these should then be combined with user models to create context-sensitive versions of the latter. The corresponding profiling techniques to be devised should be much more refined, being able to identify the relevant parts of the context at any point and interpret user actions accordingly. Such context identification will also be needed at run time so that the appropriate parts of the user's profile becomes activated. In general, contextualization and context-sensitive personalization have been studied very minimally even for the simplest of cases.

9. Multi-Modal Interaction

The previous sections have already brought this issue up. In real-life shopping, people ask direct questions to merchants, browse display windows or store shelves, are influenced by advertisements, and in general may be engaged in all these concurrently, in no particular order.

Similarly, users should be able to interact with the Open Agora in multiple ways, switching at will from one to the other, using the results of one action as input to the next independent of the interaction style used in them, and combining user-initiated (e.g., querying) and information-initiated (e.g., automatic feeds) activities. For example, while browsing some digital repositories in The Netherlands for graduate theses on Dutch folk art, Iris is notified about some forthcoming auction on Dutch and Flemish drawings. She immediately establishes a stream to retrieve every item from the auction catalog and compare it with material she already has in her personal information base. She also requests that, while examining the contents of a thesis from the repository, relevant parts of it, whether specified by Iris through some annotation or identified as important by the system, are compared against the catalog material as well as other resources in the Open Agora. Results should be displayed appropriately so that she may be able to react immediately if something significant is found.

Clearly, there are many user interface issues that arise and need to be addressed before such multi-modal interactions may be realized. In addition, several aspects of search and retrieval require rethinking to deal, for example, with modifying a query while it is being executed (e.g., adding new objects for comparison into a query comparing two collections).

10. Conclusions

In this short paper, we have introduced and discussed at some high level the concept of Open Agoras of data and information, which bring to the electronic world many concepts that are part of real-world activities in actual markets of material goods. Led by some fictitious scenario and associated examples of interactions with an Open Agora, we have identified eight key areas of concern. For each one, we have briefly described its nature and main characteristics and have outlined the most important technical challenges that it raises. We believe that the need for Open Agoras is real and current but the road to making them a reality is long. Fortunately, that road is also exciting full of research problems waiting to be explored.

Acknowledgements: We are very grateful to Fragkiskos Pentaris for his valuable contributions to several parts of this paper and his feedback on the complete text as well.

References

[1] A. K. Dey and G. D. Abowd. Towards a better understanding of context and context-awareness. In *CHI 2000 Workshop on The What, Who, Where, When, and How of Context-Awareness, The Hague, The Netherlands*, April 2000.

[2] H. Gravelle and R. Rees. *Microeconomics (3rd edition)*. Pearson Education, England, 2004.

[3] M. Machina. Choice under uncertainty: Problems solved and unsolved. *The Journal of Perspectives*, 1:121–54, 1987.

[4] J. Pitkow, et al. Personalized search. *Communications of the ACM*, 45(9):50–55, September 2002.

[5] J. S. Rosenchein and G. Zlotkin. *Rules of Encounter : designing conventions for automated negotiation among computers*. The MIT Press series in artificial intelligence, Cambridge, MA , USA, 1994.