Internet - not an Ordinary Database

Andrej Brodnik*

Abstract

In this contribution we briefly survey data manipulation over widely distributed systems such as Internet and bring forward some facts which must be observed when designing applications built based on such data. We focus on manipulation of data that have a special characteristics: the origin of data is in a small number of sites, the data does not change too frequently, and the number of clients is substantially larger than the number of data providers. Such characteristics are very common in systems which provide services for electronic commerce.

1. Introduction

Probably the two main characteristics of database systems are: to consist of two interacting agents; and they organize data appropriately for efficient manipulation. The interacting agents are: the server that is responsible to store and maintain the data, and to compute queries against these data; and a number of clients that are asking queries. Indeed, there exists another class of client agents that requests server to update the data, but for the purpose of this contribution we disregard them.

At the first moment Internet and the resources offered through it seems to fit the above description.[1] However, it fits the description only partially because the Internet lacks tools for efficient data manipulation.

Another crucial difference between database systems and Internet is in the ratio between the number of clients and servers and their spatial position. Although the number of clients in classical database systems is larger than the number of servers on the Internet clients significantly outnumber servers. This difference introduces the *asymmetry* and in the design of the data manipulation systems.

The final difference we mention is the frequency at which the data are updated. This frequency is very low, or, more precisely, usually the data are only added to the information system and almost never changed.

* IBC, Iskra Systems, Ljubljana, Slovenia and Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia. E-mail: Andrej.Brodnik@IMFM.Uni-Lj.Si.

[1] From now on we will denote with the term Internet not only the net itself, but in particular resources offered through it.

In the rest of the paper we point to some solutions found in other fields of Computer Science that solve the problems similar to ones described above.

2. Dissemination of Data

The reason to put the data on the Internet is to make them available to the broader audience; in other words we want to *disseminate the data*. The next, perhaps even more crucial step, is to make the audience *aware* of the data. There are two ways to do this: the reader can search for the data or (s)he can request to be informed when the data comes into existence. The first approach is supported through a number of searching engines (cf. Alta Vista, Open Text, Yahoo, etc.) while the second one has its roots in much older system - mailing lists [Kro92]. Though these two approaches seem to be orthogonal one to another in fact they supplement each other and can be implemented simultaneously when using appropriate techniques [BB96].

In [BB96] authors describe the KRPAN architecture which is used to implement a newspaper agency on the Internet. The data are filled in the system through the number of channels and then redistributed to subscribers. The decision which data to redistribute to which subscriber is done in two steps: first the subscriber specifies what data (s)he is interested in; later editor classifies the entering data according to a number of keys and then the data is automatically sent to the interested subscribers. Such systems are also known as *publish/subscribe* systems (cf. *SIFT - Stanford Information Filtering Tool*, [YM95, YM96]).

In traditional systems the dissemination of data is client initiated. The data is sent from the server to the client when the later requests it; we say that the data are *pulled* by the client from the server. Such an approach is optimal if either data changes very frequently and/or there are only a few requests for the same data from the same area. As shown in [CBC95] none of this is true in the systems we are considering here.

In contrast, in the publish/subscribe systems the data dissemination is server initiated; we say that the data are *pushed* from the server to the client.[2] The main drawbacks of such systems are that there are clients who might be interested in the data, but did not get it and that the system requires client to subscribe (i.e. list) to all subject (s)he is interested in. In particular the later requirement seems to post too high demand on the client.

Therefore in [BB96] is proposed a *middle* solution, where the data are still pulled by the client, but simultaneously pushed by the server as well. To implement both strategies simultaneously we introduce special purpose sites which *cache* the data. However the caching goes beyond classical caching one finds in computers (cf. proxies), as the data is pushed into caches also by the servers.[3] These sites combine properties of proxies and mirror sites (cf. [BBL96]).

---

[2] For a more complete taxonomy of push/pull systems see [FZ96].

[3] If the reader looks for the analogy in computer design, this would mean that the memory would push the data into the processor's cahce.

Such hybrid system works in two phases. First, in the pull phase, the intermediate sites build profiles of clients accessing servers. These profiles are in the push phase used by servers to disseminate the data they are offering closer to clients.

3. Practice

Although the intermediate sites need not to be country-based, in practice they should be due to topology of Internet connections. This has an additional benefit in overcoming international barriers in electronic commerce. Namely around these sites can be built teams which besides running the sites also provide services for the SME; e.g. dissemination of their data, translation of documents in other languages etc. Initially these sites would probably be founded through the (international) public funds (research and/or development) and run as part of international projects such as *EUMEDNET*. Later, however, they should be commercialized as, for example, *inReference* Inc.

4. Conclusions

In this contribution we highlighted some quite often forgotten facts about data manipulation through the Internet. Moreover, in the design of our data manipulation systems we should use the knowledge and techniques developed in related fields of Computer Science such as distributed computing and database design.

References

[BBL96] V. Batagelj, A. Brodnik and M. Lokar. Slovenian Education Network. Inter-Med 96, 1996. URL `http://www.educa.fmf.uni-lj.si/ro/izomre/novice/network.htm`.

[BB96] I. Benko and A. Brodnik. Architecture of KRPAN. Electrotechnical Review, 63(2):65-68, 1996.

[CBC95] C. Cunha, A. Bestavros and M. Crovella. Characteristics of WWW client-based traces. Technical Report TR-95-010, Boston University, CS Department, Boston, Massachussttes, 1995.

[FZ96] M. Franklin and S. Zdoni. Dissemination-Based Information Systems. Bulletin of the TC on Data Engineering, 19(3):20-30, 1996.

[Kro92] E. Krol. *The Whole Internet User's Guide & Catalog.* O'Reilley & Associates, Sebastopol, California.

[YM95] T.W. Yan and H. Garcia-Molina. SIFT - a Tool for Wide-Area Information Dissemination. In *Proceedings of USENIX Technical Conference*, 1995.

[YM96] T.W. Yan and H. Garcia-Molina. Efficient Dissemination of Information on the Internet. Bulletin of the TC on Data Engineering, 19(3):48-54, 1996.