Architecture of KRPAN*

KRPANova arhitektura

Igor Benko† Andrej Brodnik‡

April 25, 1996

Index terms: distributed computing, Internet, software architecture, ethics.

Abstract

In the paper we address the operation of an agency that provides information in a newspaper-like form on the information highway. We describe the software architecture and the physical layout of KRPAN, a kernel that provides the support necessary to operate such an agency. KRPAN is a distributed system which employs intelligent caching to improve space and network utilization. The implementation of KRPAN relies on standardized formats of data which permits usage of commonly available tools. At the end we touch legal and ethical questions and describe how KRPAN helps to solve them.

Povzetek

V članku se ukvarjamo z delovanjem organizacije, ki na informacijski prometnici posreduje informacije v obliki novic ali časopisa. V ta namen predstavimo programsko arhitekturo in fizično ureditev KRPANa, jedra, ki nudi potrebno podporo za delovanje takšne organizacije. KRPAN je porazdeljen sistem, ki uporablja inteligentno kaširanje („caching“) za izboljšanje izrabe omrežja in diskovnega prostora. Z uporabo standardnih podatkovnih formatov lahko pri izdelavi KRPANa uporabimo široko paleto in brezplačno dostopnih orodij. Na koncu članka se dotaknemo še nekaterih pravnih in etičnih problemov ter pokažemo kako nam KRPAN omogoča zadovoljivo odgovoriti na ta vprašanja.

1 Introduction and Motivation

The most popular mean of information highway (IH) browsing is WWW (cf. [2]) with a number of tools to support it (e.g. `mosaic`, `netscape`, `lynx`, etc.). The most important feature of WWW is the usage of URL (uniform resource locator) which allows simultaneous use of different services provided on Internet (e.g. `gopher`, `ftp`, `telnet`, etc. – cf. [11]). On the other hand, WWW requires an active interaction from a user in a sense that the user must explicitly connect to the source of information to retrieve it. In this contribution we study a different approach where the user provides her or his list of interests to an agent and then the agent automatically sends the requested information the user. The motivation for this approach is dual: first, the amount of new information available on IH grows so fast that it takes too much time to find it; and secondly, the user prefers to get abstracts of relevant information "at her or his fingertips" (in a mailbox) and then decide whether to retrieve the entire information or not.

`WAIS` and `archie` servers (cf. [14]) already provide a similar service by maintaining indices to existing information. These systems offer a one-time-query service, but they cannot update the user continuously in the case of changing information.

In the "paper world" a newspaper publishing agency is the agent which periodically updates subscribers with fresh information, while the one-time information retrieval is done through libraries. In the "electronic world" we have means to combine both services and

* An earlier version of this work was presented at *1995 TRIO/ITRC Research Retreat* ([1]).

†University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1. E-mail: `IBenko@UWaterloo.CA`.

‡Luleå University, S-971 87 Luleå, Sweden; on leave from Institute of Mathematics, Physics, and Mechanics, Jadranska 19, 61000 Ljubljana, Slovenia. E-mail: `Andrej.Brodnik@IMFM.Uni-Lj.SI`.

this paper suggests how to combine them using existing tools. It turns out that the model of interaction between the user and the agent can be traced back to such services as moderated mailing lists, Usenet newsgroups, DECnet NOTES etc.

The software architecture we study makes use of existent IH related tools in managing a newspaper agency. We describe "KeRnel for PAn-world Newspaper agency", or *KRPAN* for short.[1] KRPAN is not restricted to big agencies, but, on the contrary, targeted to small enterprises.

The paper is organized in the following sections: software architecture, physical layout, information formats, operation of KRPAN, and ethical questions.

2 Software Architecture

First we describe an architecture of KRPAN. As expected it is similar to the one of a "classical" newspaper agency. Roughly, we have the following processes (see Figure 1): gathering of information, editing, distribution, and technical maintenance. KRPAN has a heterogeneous architecture where the first three processes form a simple pipe-line which channels the *material information* (thick, full lines in Figure 1) from IH back to it. The maintenance process, which handles the *supporting information* (dashed thick lines in Figure 1), adds elements of a blackboard architecture (cf. [8])

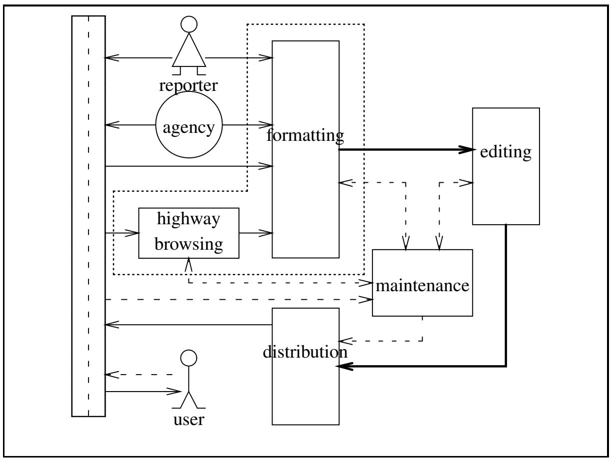

Figure 1: The architecture of KRPAN.

[1]Martin Krpan is also the name of a character from Slovene literature. In his times salt was very valuable and he was initially smuggling it on highways of Austro-Hungarian empire. However, later he got a permission from the emperor to transport salt legally ([12]).

The information gathering process in Figure 1 is split into two subprocesses: IH browsing and formatting. To describe their roles we first observe that there are three ways of information gathering: from reporters (*genuine information*), by active browsing of IH (*found information*), and from other agencies (*obtained information*). The genuine and obtained information enters into the system either directly (e.g. typing, scanning) or via the IH (e.g. e-mail). The found information, however, enters by active browsing of data repositories via IH (e.g. using `mosaic` etc.).

The found information has properties of genuine and obtained information. Similar to genuine information, we need an active "browser", a person similar to the reporter, who regularly investigates IH to obtain the new information. On the other hand, the found information is similar to the obtained information – it is not genuine and thus it is protected by the copyright of the information owner. The issue of the copyright is discussed in § 6.

The entered information is transformed by a formatting subprocess into a unique internal representation and is sent to the editing process. The editing process filters the material information and using the supporting information prepares individual issues of *bulletins* to be published. A bulletin consists of a signature and a content. The signature contains the name of the bulletin, the class of subscribers which should get the bulletin, etc.

Finally, a bulletin is sent to the distribution process which distributes it to different targets and/or subscribers on IH. The information how to do this is obtained from the supporting information and the signature of the bulletin.

The main responsibility of the maintenance process is to provide and maintain supporting information used by the other processes. Roughly, the maintenance process contains subprocesses for interaction with subscribers, for creation of signatures for new bulletins, for description of formatting patterns of out-going issues, etc. The supporting information consists of a list of possible interest classes of subscribers, of a list of subscribers, their addresses, and the interest classes they belong to, of a list of data repositories used by IH gathering process, etc. A more detail description of either the maintenance process or the supporting information is outside the scope of this paper.

3 Physical Layout

In the previous section we outlined the software architecture of KRPAN and in this section we describe its physical layout. KRPAN has an architecture with a central master copy of data and a number of distributed mirror copies (see Figure 2). The new information is added

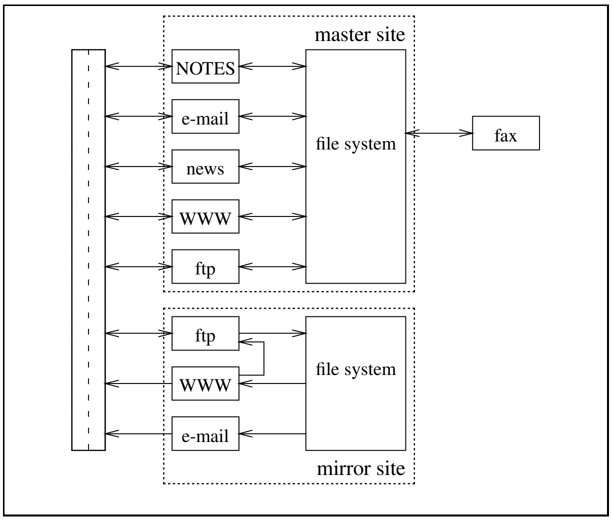

Figure 2: Physical layout of KRPAN.

into KRPAN *always* at or through the master site and subsequently distributed to the mirror sites when necessary.

The information handled by the system is either *indirect* or *direct*. The indirect information is in fact a pointer (URL) to its real location which is either at the master or at one of the mirror sites. For example, all information at a mirror site is initially indirect. However, when such information is requested, the mirror site automatically requests the missing information from the master – the information becomes direct (the mirror site behaves as an "intelligent" cache – cf. [4]). Furthermore, the direct information at a mirror site is not kept indefinitely – eventually it becomes indirect again. Finally, when a completely new piece of information is added into the system, the master site advertises its existence to all mirror sites by distributing indirect information. A more detail example is given in § 5.

As shown in Figure 2, KRPAN makes use of e-mail, Usenet newsgroups, `ftp`, and WWW. All these tools are employed to distribute information to subscribers; in addition `ftp` is used to distribute information to mirror sites. Again, a more detail example is given in § 5.

4 Information Formats

As mentioned in § 2 material information handled by KRPAN appears in many different formats. However, to freely exchange it between subprocesses of KRPAN (see thick, full lines in Figure 1) we unify its format inside KRPAN to SGML (cf. [9, 15]).

KRPAN can also infer the format of outgoing information. For example, the distribution process (see Figure 1) supports MIME and HTML formats (see [5, 3]), where the former is used for distribution of information through the e-mail and Usenet newsgroups, and the later for distribution via WWW (see Figure 2). The advantage of these formats is that they include other formats (e.g. PostScript, GIF, JPEG, etc.) and can be used to distribute an arbitrary multimedia information. However, this is not true for DECnet NOTES and non-MIME e-mail messages, which support only textual information in ASCII or ISO-8859. On the other hand, the format of fax messages, TIFF, is restricted to visual information only (see Figure 2. Of course, the incoming information can be in arbitrary format, including any of the above.

Finally, because of diversity of information formats we need powerful conversion tools: between analog and digital formats (e.g. OCR), and between different digital formats of information.[2] To describe a digital format of information we use formal grammars and to convert one into another we specify grammar conversion rules (see [13, 10]). Note that during conversion some information might get lost but there are ways to handle this problem ([6]). Nonetheless, there is another advantage to use only SGML inside KRPAN: the only conversions we have to have are between SGML and each external format which substantially reduces the number of necessary translators. The same approach was also taken by Chameleon ([13]).

5 Operation of KRPAN

In this section we illustrate the operation of KRPAN. We take a look at a life of a bulletin from the moment the information is received to the moment the bulletin is made available to subscribers.

Assume a package of articles is received from another agency or a reporter via e-mail. The articles are

[2] We use a term analog format of information to describe information which is handled as it is without understanding its content, e.g. a picture.

formatted into the internal representation (SGML). Provided that the incoming format is well defined, the formatting is done automatically, with a possible help from a system of oracles ([6]). Articles are then reviewed by an editor who selects what to distribute to individual classes of subscribers. Finally, the selected articles are formatted into the appropriate output formats and the resulting bulletins are distributed.

There are two basic ways to distribute a bulletin. One way is to place a direct information on the master site and distribute the corresponding indirect information (pointers to the direct information) among the mirror sites. When a bulletin is accessed through a mirror site, and it is present there only as an indirect information, the mirror site first downloads the bulletin and then transfers it to the user. We remark that indirect information can refer to another mirror site which may also contain indirect information. In this case the transfer of a bulletin is done in a cascading manner. We also remark that the downloading mechanism is easy to implement under WWW: indirect information invokes a script that does the downloading and finally rewrites itself with the direct information.

Another way to distribute a bulletin is to send it out by e-mail. Still, we may choose some information, like pictures, to be distributed as indirect information pointing to a mirror or the master site. It is the task of the maintenance process to define for each subscriber the site contained in the indirect information sent to her or him. The choice is based on different criteria: the proximity of a site, the reliability of a site or connection to it, the preference of a user, etc. The access to the direct information is done employing MIME standard while the transfer is performed in the same cascading manner as described in the previous paragraph.

6 Ethical questions

Because the agency handles information which is owned by somebody, this raises various ethical and legal questions. It is not our intent to discuss these questions in a great detail here, but we rather mention some of the proprietary aspects of different categories of information and suggest how to deal with them in the context of KRPAN. For more details on legal and ethical questions see, for example, [7].

First we address the status of supporting information. Our guide are newspaper agencies and not electronic mailing lists. Namely, we believe that the lists of subscribers are to be kept confidential to protect subscribers' privacy.

A more complicated issue is dealing with material information and its ownership. There is no doubt about genuine information which is a property of the agency. Also, there should be no problems with the obtained information, provided that the agreement to obtain it from another agency covers the copyright.

The most complicated is the issue of found information. To deal with it we further split it into two sub-categories: *fixed found information*, found on data repository sites of IH (e.g. WWW), and *floating found information*, found in newsgroups, mailing lists, etc. To handle the former we introduce *permanent indirect information* that is a special kind of indirect information which cannot change from indirect to direct form. Still, problems concerning redistribution of floating found information remain open, and the corresponding decisions are left to the judgment of an editor (for some guidelines see [7, pp. 47–67]).

References

[1] I. Benko and A. Brodnik. Newspaper agency on information highway. In *Proceedings 1995 TRIO/ITRC Research Retreat*, Kingston, Ontario, Canada, 1995.

[2] T. Berners-Lee, R. Cailliau, A. Luotonen H.F. Nielsen, and A. Secret. The world-wide web. *Communications of the ACM*, 37(8):76–82, 1994.

[3] T. Berners-Lee and D. Connoly. HyperText Markup Language – 2.0, November 1995. Request for Comments: 1866.

[4] T. Berners-Lee, R.T. Fielding, and H. Frystyk. Hypertext Transfer Protocol – HTTP/1.0. URL: `http://www.w3.org/pub/WWW-/Protocols/HTTP/1.0/spec.ps`, February 1996.

[5] N. Borenstein and N. Freed. MIME (Multipurpose Internet Mail Extensions), June 1992. Request for Comments: 1341.

[6] A. Brodnik and F.Wm. Tompa. A new architecture to support database updates through views. Technical Report OED-93-02, University of Waterloo, Centre for the New Oxford English Dictionary, Waterloo, Ontario, Canada, 1993.

[7] E.A. Cavazos and G. Morin. *Cyberspace and the Law: Your Rights and Duties in the On-Line World*. MIT Press, Cambridge, Massachusetts, 1994.

[8] D. Garlan and M. Shaw. An introduction to software architecture. *Advances in Software Engineering and Knowledge Engineering*, 1, 1993.

[9] C.F. Goldfarb. *The SGML Handbook*. Clarendon Press, Oxford, Great Britain, 1990.

[10] P. Kilpeläinen, G. Lindén, H. Mannila, and E. Nikunen. A structured document database system. In *Proceedings of the International Conference on Electronic Publishing, Document Manipulation & Typography*, Gaithersburg, Maryland, 1990.

[11] E. Krol. *The Whole Internet: User's Guide & Catalog*. O'Reilly, Sebastopol, California, 1st ed., minor corrections edition, 1993.

[12] F. Levstik. *Martin Krpan*. Mladinska knjiga, Ljubljana, Slovenia, 1994. (In Slovene.).

[13] S.A. Mamrak, J. Barnes, and C.S. O'Connell. Benefits of automating data translation. *IEEE Software*, 10(4):82–88, July 1993.

[14] K. Obraczka, P.B. Danzig, and S.-H. Li. Internet resource discovery service. *IEEE Computer*, 26(9):8–22, 1993.

[15] Joan M. Smith. *SGML and Related Standards*. Computers and Their Applications. Ellis Horwood, New York, 1992.

Igor Benko received the Dipl.Ing. degree in computer science from University of Ljubljana in 1991, and a MMATH degree in computer science from University of Waterloo in 1993. Currently he is a PhD candidate in the Department of Computer Science at University of Waterloo. He is on the leave from the Department of digital communications and networks at the Jozef Stefan Institute, Ljubljana, Slovenia. His research interests include designing concurrent systems, in particular asynchronous circuits design, silicon compilation, and formal specification techniques.

Andrej Brodnik received Dipl.Ing and MSc in computer science in 1984 and 1988 respectively from the University of Ljubljana, Slovenia, and PhD in computer science from the University of Waterloo, Ontario, Canada. Currently he is a postdoc at the Luleå University, Sweden. His research interests include data structures, algorithms, formal languages, and Internet technologies.