Automated organization of caches architecture

Luigi Lancieri France Telecom CNET
luigi.lancieri@cnet.francetelecom.fr
WebNet Conference 1998

1. Introduction

We consider the case of a multi-discipline company distributed in several sites. These sites are linked with a sufficiently dimensioned Intranet which allows a good quality of service. We use caches to increase the comfort of the users for an external connectivity to Internet. These caches are organized so as to cooperate and so to optimize performances. The objective of this proposal is to use the principle of the learning process to direct automatically each user towards the most liable cache to contain his requested page. Furthermore, we describe, according to the same principle, an automatic organization method of cooperating caches architecture. In this context, it is more interesting to use several thematic self dedicated caches instead of a « big central» one. Each dedicated cache will correspond to a specific activity of the company.

2. General presentation

In this document, the expression « local cache » makes reference to a shared cache which is local to a site and not to a personal browser. Cooperating caches architectures are now rather spread, figure 1 shows us an example. In such an architecture, the users set up their browser to get connected to the nearest first level cache. This one questions the second level cache when it can not find the requested page in its memory. Finally, if the third level cache does not have the page, it questions directly the distant server. This kind of architecture is rather common and works with the HTTP protocol for the transport of the main data and the ICP protocol for the communication between caches. It allows good performances comparatively to a unique central cache. The following part of this document presents some ideas in order to optimize the performances of cooperating architecture. In our proposal, users will be automatically grouped in several virtual communities of interests. We introduce the dynamic organization principle based on the thematic profile of the cache content. In one word, each user will be directed towards the first level cache that contains data closed to its center of interest. In this condition, the probability to find the requested page (Hit rate) is higher than in the regular architecture. Figure 2 highlights the most dynamic aspect of this configuration.

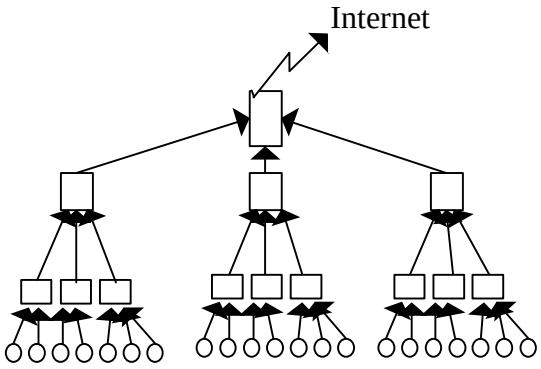

Figure 1. Regular cooperating architecture

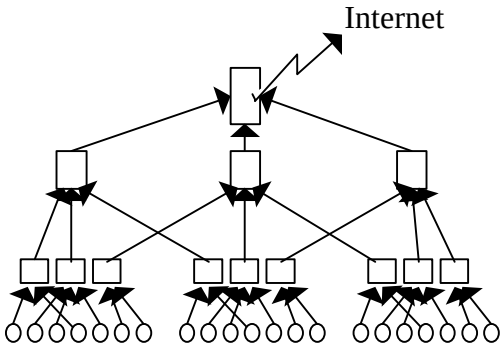

Figure 2. Dynamic cooperating architecture

This organization is based on the observation and the analysis of the users' connectivities. In the initial situation, the users are directed towards the nearest first level without other consideration. In the same way, each first level cache is directed towards the nearest second level . The second phase described below consists in identifying the different users profiles and in categorizing these profiles in a limited group of topics corresponding to the number of the first caches level.

3. Identifying users' profile

The analysis of the cache content as well as the caches log files gives us precious information about the users' behavior. Indeed, the cache contains pages and all the necessary information which allows to identify the pages

viewed by each user. A specific algebraic model of semantic association [1] based on neural networks allows a self-extraction of themes corresponding to a set of multimedia documents. Operating this model on the cache content allows to extract the center of interest of the users. In this model, the cache is used as an « implicit filter ». This means that it will only contain, according to its operate mode, data corresponding to the profile of users who are connected. The lexical analysis of the cache content allows to build a semantic network where each neuron represents words. After the learning period, the weight of the links between neurons will be as high as the semantic links between associated words. The algebraic representation of this semantic network allows to extract interesting properties using regular mathematic tools. In this representation, each word is associated to a vector in a global matrix so it is possible to evaluate « semantic distances » by calculating Euclidean distances between vectors. It is also possible by calculating proper vectors of the global matrix in order to bring out the most representative words from the analyzed documents. These words are those to which other words make the most often reference directly or not. We identify these words as themes. In the same way, it is possible to extract themes in a sub set of pages only downloaded by a specific user or a delimited group of users. We identify these specific themes such as the users' profile. Since a theme and profiles are represented by vectors, it is convenient to evaluate if two profiles are closed to each other by calculating the Euclidean distance. All the details about this aspect of the paper can be find in a referenced paper.

4. Interconnection between users and first level caches

So, the goal is to group the users who have common interests towards a specific first cache level . As we explained previously, the first step is to identify the users' profile, then to categorize them, and finally to allocate a theme or closed group of themes to each first cache level. It is now necessary to connect dynamically users to their « dedicated » first cache level. This is done through the auto configuration file (.cap) capabilities of the web browser. The principle of such a configuration is the following : the user indicates one and for all during the configuration of his browser the path to a server that contains the auto configuration file. This file is actually a javascript automatically downloaded by the browser which describes the rules of the connection between users and caches. The administrator of the network have the ability to change the content of this file and so the rules of connectivity. In our case, the management of this file is automatically operated by a process that takes into consideration users' profiles. It is indeed important to see that the connectivity rules move during the time because they are based on a learning process which is fundamentally dynamic.

5. Interconnection between caches

This part describes the automatic configuration of the connection between first and second caches level. The goal was first to determine the profile of each user, we have now to determine the profile of each first level cache. This could be considered superfluous at first because we already allocated themes to these caches. Actually, we have to consider that this allocation was based on an approximate analysis grouping in one component several users. These users have a fluctuating behavior including « noise » (connection without relation with the allocated topic). So, to « integrate » the behavior of several users it is necessary to analyze the whole cache content. So, once we have done this for all first level caches, we have to categorize all these first level caches profiles regarding to the amount of second caches level. Finally all second level caches will be directed towards the unique third level cache.

6. Conclusion

The main idea of this proposal is to automatically identify groups or communities of interests on an Intranet. The organization of caches' architecture according to these groups of interest allows a higher global performance. This ability also allows to envisage other form of services such as the automatic diffusion of information according to users' profiles or the control of access to documents for example.

7. References

[1] Distributed Multimedia Document Modeling

Luigi Lancieri IJCNN 98 (IEEE Joint Conference on Neural Networks)