# Collaborative Document Edition in a Highly Concurrent Environment

Willy Picard
*Department of Information Technology*
*The Poznań University of Economics*
*Mansfelda 4, 60-854 Poznań, Poland*
*picard@kti.ae.poznan.pl*

## Abstract

*Up to date, research on computer-supported cooperative work (CSCW) focused mainly on the adaptation of collaboration models used in Intranet to the specificity of Internet. No collaboration model deals with the problem of high amount of data resulting from the high number of collaborators. In this paper, a model of document edition in highly concurrent environments is presented. A document model based on multiversion databases is introduced. An overview of ultra-metrics and hierarchical classification is given, and the use of these techniques for edition process analyses is presented. The use of ultra-metrics in the edition process is discussed, both for simulations and automatic edition steering. Three examples of applications of the proposed document edition model are given.*

## 1. Introduction

From prehistoric tribes to trade unions, group structure has always been at the heart of human activities. Grouping their competences, humans are able to achieve great projects, from pyramids to railroad infrastructure construction. The key word for group activities is *collaboration*. Collaboration is the process of sharing competences to achieve a common goal.

To a recent past, the collaboration process was limited by the requirement of a single location. People involved in a collaboration process needed to meet to exchange information. In reality, people are generally spread on large geographical area. Meetings are difficult to organize, because of schedule incompatibilities, and costly in terms of time and money.

Telecommunication networks provide a partial solution to the above problem. Telecommunication networks let collaborators be spread over various locations. The use of telephone allows collaborators to exchange information via voice communication. Documents can be exchanged via fax in a graphical format. Local area networks (LAN) are the basis of electronic information exchange inside enterprises, while wide area networks (WAN) - in between different enterprises.

With the rise of telecommunication networks, collaboration models that rationalize the collaboration process have been developed. Most of them are document oriented, i.e. the fundamental object of the collaboration process is one or more documents. In companies, collaboration tools are currently widely used for sharing files, for group scheduling or for document collaborative writing.

Usually, telecommunication networks can connect only a limited number of users at a given time: LANs connect people inside a given company, telephone is generally used for communication between two persons, sometimes for phone conferences with up to ten participants. On the opposite, Internet allows an unlimited number of persons from the whole world to collaborate. As a consequence, most works in the collaboration research field focus nowadays on the adaptation of collaboration models developed originally for intranets to the needs of Internet. Classical research areas are shared environments, collaborative drawing and writing, and workflows. Shared environment applications [1] aim at providing a virtual common place for collaborators. They are generally divided in applications based on virtual reality and applications aiming at window and service sharing. Collaborative drawing and writing applications [2] allow multiple users to simultaneously work on a given document or drawing. In the workflow field [3], works focus on managing, documenting, automating, and (if necessary) reengineering business processes and workflows, enabling organizations to be more efficient and agile. They focus on the problem of systematically coordinating small process models to achieve coherent business operations.

All these models address the issues related to the adaptation of known models to the needs of Internet. However, none of them addresses the problem of the potentially high amount of data resulting from the collaboration process in the Internet. The amount of data exchanged during the collaboration process is high. This

important amount of data can follow from a high number of collaborators. The higher the number of collaborators, the greater amount of data resulting from the collaboration process. Another cause of the important amount of data is the versioning requirement. The collaboration process must be archived that means that the old data must be stored. As a result, the amount data generated during the collaboration process growths with such addition of new versions. Human beings cannot deal with this great amount of data without analyzing tools that can provide a synthetic view of the collaboration process status.

In this paper, we propose a document edition model that addresses the problem of scalability in highly concurrent environment. In Section 2, a document model that deals with the issues related to the high number of collaborators in highly concurrent environment is presented. In Section 3, an edition analysis mechanism based on ultra-metrics is discussed. In Section 4, two solutions to the problem of document convergence are given. In Section 5, three applications of the presented edition model are presented. Section 6 concludes the paper.

## 2. Document model

Document edition is a classical topics of collaboration research. Many edition models have been presented. They can be grouped in three generally accepted edition models: serialized, centralized and parallel.

In the serialized model, it is assumed that only one collaborator is modifying the document at a given time. In this model, a collaborator modifies the document, which is then accessible for the other collaborators in the read-only mode. When all modifications are provided, the collaborator frees the document that can then be edited by another collaborator. A known implementation of this model is the cooperative edition module of the Microsoft Word editor. In the case of highly concurrent environments, this model presents high latency due to the important number of collaborators waiting for the document.

In the centralized model, many collaborators may edit a document at a given time, assuming that only some parts of the document are locked by some collaborators. If a collaborator starts the edition of a document part, other collaborators can have access to this part in the read-only mode. However, any collaborators may edit any part that is not locked. In the case of highly concurrent environments, the number of collaborators is generally much greater than the number of parts. When all parts are locked by collaborators, the remaining collaborators are waiting for parts to be freed. Moreover, in document edition, all parts of the document are not of equal interest for collaborators. Generally, only a few parts are modified and these parts are the bottleneck. This model induces some important latency of document edition in highly concurrent environments.

In the parallel model, it is assumed that every collaborator has its own document version. Each collaborator is modifying its own version independently of others collaborators. There is no bottleneck. The latency of this model is not an issue. On the contrary, the main issue to deal with is the merging process. Automatic merging is not yet solved and human intervention is always required.

In the case of document edition in highly concurrent environments, the only acceptable model is the parallel one. However, the document model must address the merging problem in order to ease the document edition. The merging problem is usually addressed by document comparison. If documents are identical or the differences occur in different parts of the document, no conflicts occurs and the merge process can be automated, on the condition that merging rules are established. If the differences cannot be removed by the merging system, human intervention is required. In theses merging systems, the merging process is seen as a side effect of the edition and is not integrated in the edition process.

The proposed document model integrates the merging process, which occurs during the whole time of document edition. Indeed, the merging process is the key element of the document edition process. Collaborative document edition means merging collaborators' preferences.

The document model proposed in this paper is based on the multiversion database model, presented in [16]. In the proposed document model, the document is a tree of document versions and consists of multiversioned parts. For each multiversion part, an *association table* maintains a mapping between part versions and document versions. These tables simplify modifications of the document and allow easy differencing of various document versions. With this model, the full tracability of the document evolution in time and among the various collaborators is achieved at the part level. The merging process can take advantage of this tracability in order to detect conflicts between different document versions. Two document versions are identical if all document parts are identical, which means that association tables of document parts associate the same part version to document versions. As the tracability is achieved at the part level, no additional text parsing is needed during conflict detection and thus performance is improved.

A major advantage of the multiversion database model is the orthogonality of versioning and concurrent accesses. In the context of document edition, this characteristics of the multiversion database model allows many collaborators to edit various versions of the

document at the same time in a very efficient way. The multiversion database model is well adapted to highly concurrent environments.

The document can be formalized as a set of points in a three dimensional space, called *D*. The space *D* is defined as $P \times V \times U$, where *P* is the document part space, *V* is the document version space and *U* is the users (collaborators) space. In Figure 1, a graphical representation of this model is given: two collaborators $U_1$ and $U_2$ are co-writing a document composed of two parts ($P_1$ and $P_2$). Collaborator $U_1$ has two versions of part $P_1$.

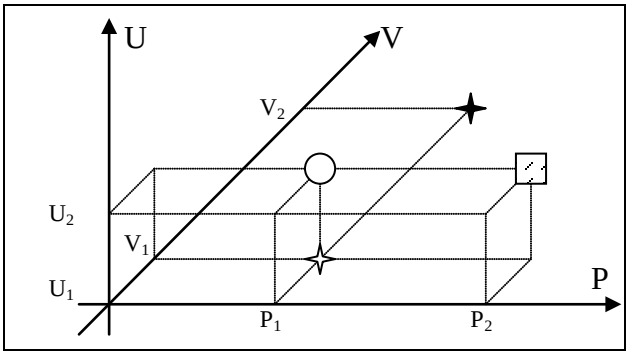



**Figure 1. Graphical representation of space *D*.**

## 3. Edition process analysis and ultra-metrics

In highly concurrent environments, it is difficult to deal with a great number of collaborators and document versions. Collaborators can be easily lost in the great amount of data. Analysis of the edition process is fundamental in all edition processes, but, in the case of highly concurrent environments, the analysis must be computed. In the document model presented in this paper, the edition process can be automatically analyzed.

The analysis of the edition process is based on the automatic classification theory and the use of ultra-metrics [17][18][19].

The mathematical definition of ultra-metrics is the following:

$$d \text{ is an ultra - metrics on space } D$$

$$\Leftrightarrow$$

$$\begin{cases} d : D^2 \to \Re^+ \\ \forall (x,y) \in D^2, d(x,y) = 0 \Leftrightarrow x = y \\ \forall (x,y) \in D^2, d(x,y) = d(y,x) \\ \forall (x,y,z) \in D^3, d(x,y) \le \max(d(x,z), d(y,z)) \end{cases}$$

Having such an ultra-metrics, the automatic classification theory guarantees that space *S* can be partitioned into classes. Moreover, this classification is hierarchical, which allows controlling the space clustering granularity.

In the proposed model, ultra-metrics and automatic classification are used to analyze the edition process. Edition analyses can be used to detect anomalies in the edition process. Some ultra-metrics can be used to detect collaborators that do not work on the document or that systematically reintroduce old versions of the document. Other ultra-metrics can detect collaborators having an incoherent behavior, proposing for instance illogical data with huge variations between various document versions. In the case of supervised editions, this data may indicate that some changes are needed, or that a given collaborator should leave the edition process.

In the edition analysis model, it is assumed that many analyses can be computed and do not impose any restriction on the choice of ultra-metrics. The only requirement for all analyses is the existence of a function *f* from *D* to the given ultra-metrics domain, called $D_D$. The function *f* preprocesses the document space *D* to provide views of the edition process that are further analyzed with the help of ultra-metrics. With the two-phase analysis, it is possible to analyze facts that are not directly accessible from space *D*, as for instance collaborators' activity rate. In this case, the function *f* processes the activity rates from space *D*, and the results are analyzed by an ultra-metrics to provide hierarchical classification.

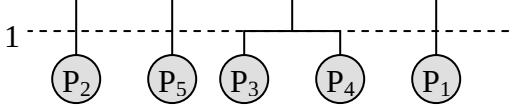
To illustrate the above technique, we present an example of the whole edition process analysis. In this example, we want to evaluate the weight of the various document parts in the edition process. We assume that the more a given part was modified, the higher the interest of this part is. Function *f* is used to generate space $D_D$ that consists of multiversion document parts. Each element of $D_D$ consists of all versions of a given document part. For example, card($D_D$)=5 and $P_1$, $P_2$, … $C_5$ are the various multiversion document parts. The values of the ultra-metrics *d* on $D_D^2$ are given in Figure 2.

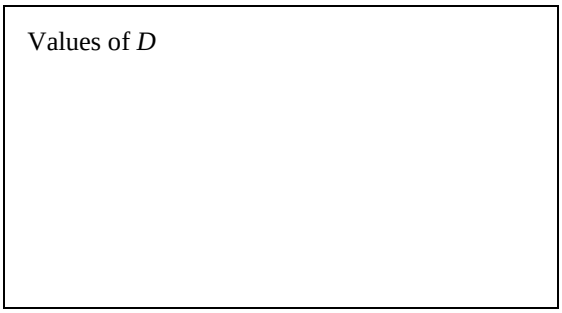
| *x* | *y* | *d(x,y)* |
|---|---|---|
| $P_1$ | $P_2$ | 7 |
| $P_1$ | $P_3$ | 5 |
| $P_1$ | $P_4$ | 5 |
| $P_1$ | $P_5$ | 7 |
| $P_2$ | $P_3$ | 7 |
| $P_2$ | $P_4$ | 7 |
| $P_2$ | $P_5$ | 3 |
| $P_3$ | $P_4$ | 1 |
| $P_3$ | $P_5$ | 7 |
| $P_4$ | $P_5$ | 7 |

**Figure 2. Values of *D* on $D_D^2$**

In Figure 2, the values of ultra-metrics *d* are presented. The hierarchical classification of $D_D$ built on the basis of theses value is presented in Figure 3.

**Figure 3. Classification of $D_D$ according to *d***




This hierarchical classification can be seen at various detail levels. Having a given threshold $T$, the space $D_D$ can be partitioned into different classes. In Figure 4 (resp. Figure 5), the partition obtained with threshold $T$=4 (resp. $T$=6) is presented.

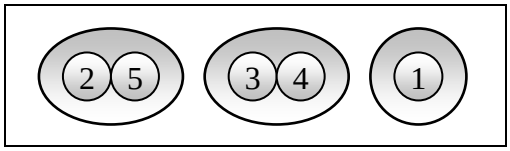



**Figure 4. Partition of space $D_D$ with threshold $T$=4**

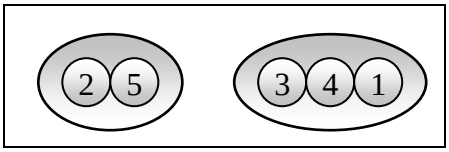



**Figure 5. Partition of space $D_D$ with threshold $T$=6**

Changing the threshold, it is possible to have control on the granularity of space partition. In the context of highly concurrent environments, this ability enables collaborators to evaluate efficiently the status of the edition process. A high threshold gives a high level classification (with a few classes) while a low threshold allows fine-grained classification (with many classes).

The choice of the threshold and the possible use of many ultra-metrics provide a very flexible framework for document edition analyses. The capability to analyze every aspect of the edition process combined with the hierarchical classification allows to focus on a given problem and to view the results of the analyses at various detail levels.

## 4. Convergence strategies

The parallel approach to the document edition imposes a new model for the edition process. This new edition model is based on the fact that every document edition process tends to reduce the differences among collaborators' document versions in order to converge to an agreement. As a direct consequence, the differences between various document versions must be measurable. The analysis of the edition process is thus the basement of the proposed model. Having such an analysis mechanism, strategies must be built to reduce the differences between document versions.

The first strategy consists of automatic confrontation of collaborators. On the basis of edition process analysis, collaborators are put in contact, according to a given matching strategy. The strategy matches collaborators in such a way that document versions converge to a common agreement. An example of such a matching strategy may consist in matching collaborators whose document versions are the most distant. Such a matching strategy ensures the reduction of the overall document dispersion, and helps document versions convergence.

The differences between documents can also be reduced by the use of edition process analyses as a simulation tool. A collaborator can evaluate the consequences of a given modification of her/his document on the whole edition process. With such an evaluation mechanism, the collaborator can decide if the given modification goes toward agreement i.e., if the differences between document versions will reduce. This simulation mechanism allows collaborators to have a better readability of the edition process and to better estimate the consequences of their own modifications.

## 5. Examples of applications

Many applications to the proposed model are possible. Different kinds of documents induce different applications. In this section, three application examples are described. Each of these examples are based on a different document category.

The first application example deals with documents defining standards. Standards are documents defining a set of agreements on procedures, data and/or formats. The standardization process is usually costly because of necessary face-to-face meetings. Moreover, in large standardization process (such as MPEG [20]), a few hundreds of collaborators are involved in the document edition process. Many versions of the standard exist and are improved by collaborators' suggestions, proposals or remarks. It is currently very difficult to have an overview of the standardization process.

The proposed model proposes a solution to the listed above issues. The multiversion database model and its application to the document model allows a full tracability of the standardization process. The flexibility of the two-phase analysis allows the processing of various analyses of the standardization process. Automatic hierarchical classification allows collaborators to focus on the standard part that really interested them. Classification

may also be used to have an overview of the standardization process, which is currently a costly task requiring the redaction of summaries, minutes, etc.

The second application example deals with software development. Documents involved in the development process are various and we focus only on source code files. In the Open Source (OS) movement, software is generally developed by many developers spread all around the world. For instance, the Apache web server has been developed by hundreds of developers with the help of CVS software and mailing-list. The CVS tool suite cannot currently present synthetic view of the development process. Moreover, in the OS movement, only some persons can update the software in order to maintain control over the development process. Such a moderated development induces additional costs.

The proposed model can significantly fasten the development processes in the OS movement. People who wants to involve in the development process can use the analyze mechanism to understand faster the whole program. Moreover, modification tracability, which is a classical issue in OS development, can be simplified by the use of ultra-metrics. In some developer in involved in some part of the program, let's say the graphical user interface, she/he can perform analyses of this program part to identify who are the contributors to this part, to find incoherencies between version, etc.

The third application example deals with contracts. Contracts are document defining agreements between contractors on various goods and clauses. In this example, we assume that a travel agency wants to build a new catalog for its customers. Each offer to the final client consists of many contracts with subcontractors. The travel agency needs to negotiate many contracts with many hotels, car-rentals, transport companies, etc.

With the help of the proposed model, the agency can build a contract that consists of many subcontracts and control the negotiation status with hierarchical classification. Ultra-metrics can also be used to compare offers from similar subcontractors, offers that generally are differently formatted. The travel agency can then compare two hotels in Tunisia, one proposing excursions in the desert and the second being in the center of the capital. The travel agency can also use the classification tool to detect and check dependencies between subcontracts. For instance, hotel reservation does not make sense if no transportation exists at the given period of time, and vice-versa.

## 6. Conclusions

The great amount about of data resulting from the collaboration process in highly concurrent environments can be analyzed by the tools proposed in this paper. The proposed model ensures that collaborators can have various synthetic views of the document edition process. Collaborative edition of complex documents can be performed by a high number of users in an efficient way, opening doors to new collaborative applications.

A prototype of the presented model is currently under development. XML is used to describe documents. The multiversion database model, which is the fundament of the prototype, is implemented in the open source dbXML-Core. The choice of an XML database has been done due to the mismatch of XML structure (tree-structure) with relational database structure (tabular data). Future research needs to be conducted to take into account mathematical properties of quantifiable data such as prices or quantity. With these data, the existence of center of mass may allow more precise analyses. Another field for further works is the integration of meta-data about negotiation with the negotiation model.

## References


[1] R. Reinema, D. A. Tietze, R. Steinmetz, "IMCE - An Integrated Multimedia Collaboration Environment", *Poster Proceedings of the Sixth ACM InternationalMultimedia Conference (MULTIMEDIA'98)*, Bristol, 1998

[2] M. Koch, J. Koch, "Application of Frameworks in Groupware - The Iris Group Editor Environment", *ACM Computing Surveys Syposium on Frameworks*, 1998

[3] A. Lazcano, G. Alonso, H. Schuldt, C. Schuler, "The WISE approach to Electronic Commerce", *International Journal of Computer Systems Science & Engineering*, Sep 2000.

[4] W. Cellary, G. Jomier, "Consistency of Versions in Object-Oriented Databases", *Proceedings of the 16th VLDB Conference*, Brisbane, Australia, August 1990.

[5] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, New-York, Wiley, 1973.

[6] B. Dubuisson, Diagnostic et reconnaissance de formes, Hermès, Paris, 1990.

[7] I. Block, "Classification Automatique", *Bases de la reconnaissance des formes (Chap. 4)*, ENST. Paris, 1997.

[8] Moving Picture Experts Group (MPEG), http://www.cselt.it/mpeg/