Université de Montréal

Extraction de motifs dans la rédaction collaborative sur les Wikis

par
Jeanne d'Arc Uwatowenimana

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en informatique

Juin, 2008

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Extraction de motifs dans la rédaction collaborative sur les Wikis

présenté par :

Jeanne d'Arc Uwatowenimana

a été évalué par un jury composé des personnes suivantes :

Guy Lapalme
(président-rapporteur)

Petko Valtchev
(directeur de recherche)

Esma Aïmeur
(codirectrice)

Yann-Gaël Guéhéneuc
(membre du jury)

Mémoire accepté le : .....

RÉSUMÉ

Depuis un certain temps, le monde de l'Internet est marqué par une façon originale de collaborer : les Wikis ou sites Web permettant aux internautes de participer facilement à la rédaction de leurs contenus. De nos jours, le Wiki le plus connu est Wikipedia, une encyclopédie libre, multilingue, écrite de façon collaborative par des volontaires.

Comment expliquer qu'un ensemble d'individus qui ne se connaissent pas nécessairement et qui travaillent dans un environnement ouvert à tout le monde, parviennent à rédiger ensemble un document tel qu'un article d'encyclopédie ? Quelles sont les interventions effectuées par ces personnes et comment s'enchaînent-elles au cours d'une collaboration de façon à assurer la qualité du document ?

Dans ce mémoire, nous mettons en évidence les manifestations des mécanismes de collaboration, tels que le maintien de la cohérence et la recherche d'un consensus, qui émergent d'une rédaction collaborative des contenus par les Wikistes.

Nous utilisons particulièrement l'*extraction des motifs séquentiels*, une technique de la *fouille de données* pour extraire des séquences d'interventions répétitives à partir des *historiques des modifications* ou *journaux* des pages. Ces derniers documentent les changements apportés par les participants dans le processus de rédaction des contenus sur un Wiki.

Dans un premier temps, nous allons constituer une classification (ou *taxonomie*) des interventions basée sur leurs aspects *syntaxique* (ou type de modification) et *sémantique*. Dans un second temps, nous allons identifier des interventions à partir des journaux de Wikipedia et les organiser dans des séquences. Ces dernières vont par la suite être annotées en utilisant la taxonomie des interventions puis être analysées afin d'en extraire des séquences (de catégories) d'interventions répétitives.

Mots clés : Wiki - Historique des modifications - Taxonomie d'interventions - Motifs de collaboration - Relations sémantiques - Fouille de données.

ABSTRACT

Recently, an original way of collaboration has appeared on the Web: a Wiki, an Internet site that allows everyone to read, modify and contribute to its contents. Nowadays, the most popular Wiki is Wikipedia, a free content and multilingual encyclopedia that is written collaboratively by volunteers.

How to explain the fact that individuals who do not necessarily know each other, working in an environment free of access by anyone and without rules, are able to write collaboratively a document such as an encyclopaedic article? What are the diverse types of interventions performed by those individuals? How to order those interventions during the collaborative process to keep the quality of a document?

In this study, we underline the manifestations of new collaboration mechanisms emerging during a collaborative redaction by people who write contents on a Wiki Website. We particularly use *Mining Sequential Patterns*, a *Data mining* technique for identifying repetitive sequences of interventions from *editing history*. An editing history contains information about users' interventions during the redaction process of contents in a Wiki Website.

Firstly, we have considered user-interventions, by focusing on both syntactic (or modification operations) and semantic (or meaning relatively to previous interventions) aspects of those interventions, we have developed a *taxonomy* of user-interventions. Secondly, we have organized users' interventions of a sample of Wikipedia articles in structures that explicit their temporal aspect. Then, those structures were annotated with categories (from the taxonomy) previously identified and analyzed to extract *patterns* or repetitive sequences of user-interventions.

Keywords Wiki - Editing history - Taxonomy of user-interventions - Collaborative patterns - Semantic relations - Data mining.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

REMERCIEMENTS

Je tiens tout d'abord à remercier le *Centre de recherche interdisciplinaire sur les technologies émergentes*, le CITÉ de l'Université de Montréal pour le financement de ce projet.

Un grand merci à toutes les personnes qui ont participé de près ou de loin à la bonne réalisation de ce travail.

Monsieur Petko Valtchev, professeur à l'Université de Montréal et directeur de mémoire pour son soutien, ses nombreux conseils et sa patience tout au long de cette maîtrise.

Madame Esma Aïmeur, professeure à l'Université de Montréal et codirectrice de mémoire pour ses critiques et ses remarques sur le travail effectué.

Madame Chantal Benoit-Barné et Monsieur Nicholas Bencherki, chercheurs au Département de communication de l'Université de Montréal pour leur collaboration.

Monsieur Daniel Memmi, professeur à l'Université du Québec à Montréal pour ses remarques et ses conseils concernant ce travail.

Je remercie également Messieurs Guy Lapalme et Yann-Gaël Guéhéneuc, professeurs à l'Université de Montréal, pour leurs précieuses remarques.

Finalement, un merci tout particulier à mon conjoint Alexandre Hugo pour son soutient de tous les jours.

CHAPITRE 1

INTRODUCTION

Notre objet d'étude, un Wiki, est un site Web qui permet à tout utilisateur de lire, mais aussi de modifier et d'enrichir directement son contenu en ligne. C'est un concept qui a été introduit par Ward Cunningham en 1995 et utilisé depuis l'année 2000 dans plusieurs domaines. De nos jours, le Wiki le plus connu est Wikipedia, une encyclopédie libre, universelle, multilingue et écrite de manière collaborative sur l'Internet.

Les informations publiées sur les Wikis sont fréquemment mises à jour par des groupes d'individus. C'est cet aspect communautaire et collaboratif qui fait la spécificité des Wikis. Cependant les mécanismes de collaboration, comme le maintien de la cohérence textuelle et la recherche d'un consensus, appliqués par les participants sur les Wikis sont encore mal compris.

Dans le cadre d'un projet mené conjointement avec les chercheurs du département de communication de l'Université de Montréal, nous avons élaboré une méthodologie de mise en évidence des manifestations de ces mécanismes à partir des journaux ou logs des pages d'un Wiki (en particulier Wikipedia). Ces journaux ou historiques des modifications documentent les changements apportés par divers participants dans le processus de rédaction collaborative des contenus.

À cet effet, nous avons explicité les rapports entre les interventions individuelles des wikistes que nous avons analysées afin d'en extraire des régularités. Ces dernières sont par la suite présentées à un spécialiste en communication qui les interprète afin d'en dégager les mécanismes effectifs de collaboration propres aux Wikis.

1.1 Site Web Wiki

"Un site Web Wiki[1]*, est un site Web* ***collaboratif*** *où chaque internaute visiteur peut participer facilement à la rédaction de son contenu."* [6]. C'est une technologie récente qui a été introduite par un informaticien développeur américain Ward Cunningham. La première implémentation d'un tel site[2] pour organiser le contenu du site du Portland Pattern Repository fut le 23 mars 1995 [24].

L'idée d'un site Wiki a réellement pris son essor cinq ans après son introduction. Ainsi, ce n'est qu'au début de l'année 2000 que plusieurs sites Wikis ont été développés et utilisés dans plusieurs domaines (en informatique : clubic, en médecine : Wikihealth, en droit : Jurispedia, dans l'industrie : Google, Toyota, etc. ou encore dans certains établissements scolaires comme l'Université de Calgary). C'est aussi durant cette même période que les "moteurs de Wiki" (ou logiciel permettant de réaliser des sites internet basés sur le principe d'un Wiki) sont apparu (exemple : Mediawiki, SocialText, jotspot, etc.).

De nos jours, le Wiki le plus connu et important est Wikipedia, une encyclopédie libre, universelle, multilingue, écrite de façon collaborative sur l'Internet avec la technologie Wiki. En mars 2008, Wikipedia a atteint la barre des 10 millions d'articles écrits en plus de 250 langues dont le deux millionième article pour sa version anglaise a été ébauché le 10 septembre 2007.

Wikipedia est un projet qui a été introduit par Jimmy Wales et Larry Sanger, le 15 janvier 2001 lorsque la version anglaise de cet encyclopédie a vu le jour. D'après le site Web Alexa[3] qui fournit des statistiques sur le trafic Web des sites indexés, Wikipedia fait parti du top 10 des sites Web les plus visités dans le monde (à la 9ème position au 1 avril 2008).

---

[1] Le mot *Wiki* vient de l'hawaïen *wikiwiki* qui signifie "vite" [6]
[2] http ://c2.com/cgi/wiki ?FrontPage
[3] http ://www.alexa.com/

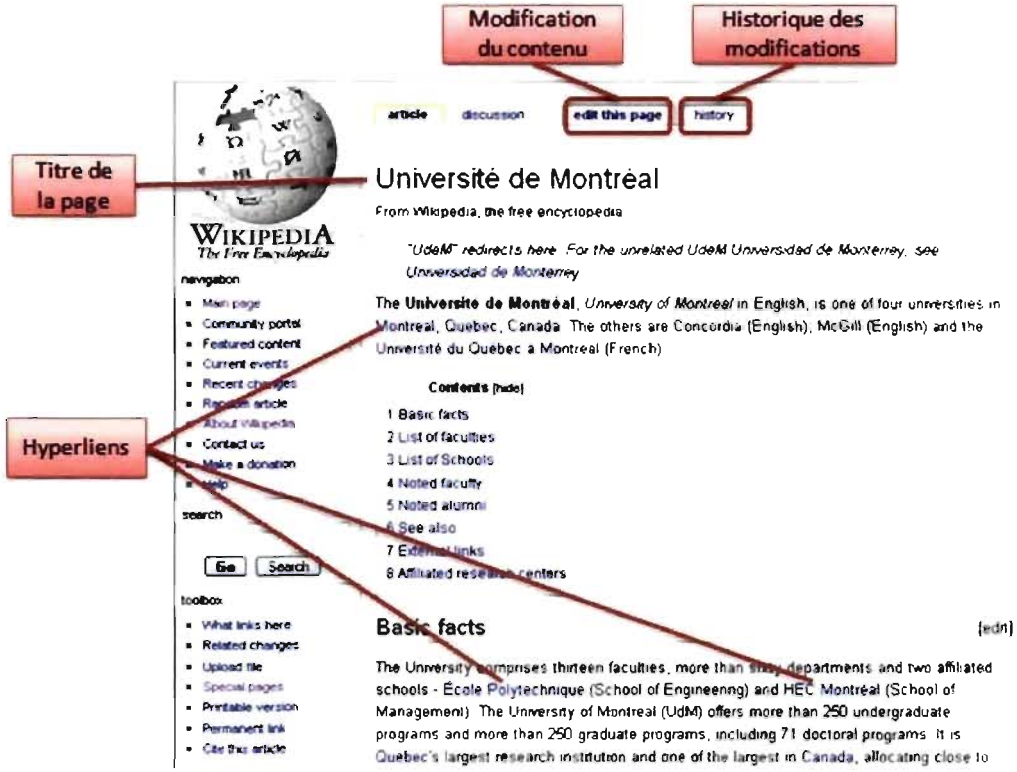

Figure 1.1 – Page "Université de Montréal" - Wikipedia

La structure générale de Wikipedia (et d'un site Wiki en général) repose entièrement sur les pages et les liens entre celles-ci (appelés également liens internes). Il existe plusieurs types de pages, néanmoins le type de pages le plus important est la page *principale* (dite *Article* dans Wikipedia). Cette dernière est consacrée à un sujet bien précis et les autres types de pages comme l'historique, la page des modifications viennent s'y attacher. La figure 1.1 à la page 3 donne un aperçu de l'article "Université de Montréal" de Wikipedia.

Chaque page principale d'un Wiki présente quatre caractéristiques principales (identifiées en rouge sur la figure 1.1 à la page 3) :

- le titre (le sujet) ;
- le lien "edit this page" qui permet aux utilisateurs d'accéder à un éditeur de texte et de modifier le contenu de cette page principale ;

– le lien "history" qui permet d'accéder à l'historique des modifications. Ce dernier est en quelque sorte un fichier log qui décrit toutes les actions qui se sont déroulées sur cette page principale. Ce journal comprend une liste de toutes les versions de la page principale avec une fonctionnalité *"diff"*, permettant de visualiser la différence entre deux versions successives (ou l'intervention de l'auteur). En plus du texte de chaque version, on y trouve le nom (alias) ou l'adresse IP de son auteur et la date de sa création. L'image qui suit montre une partie de l'historique des modifications correspondant à l'article "Université de Montréal" de Wikipedia ;

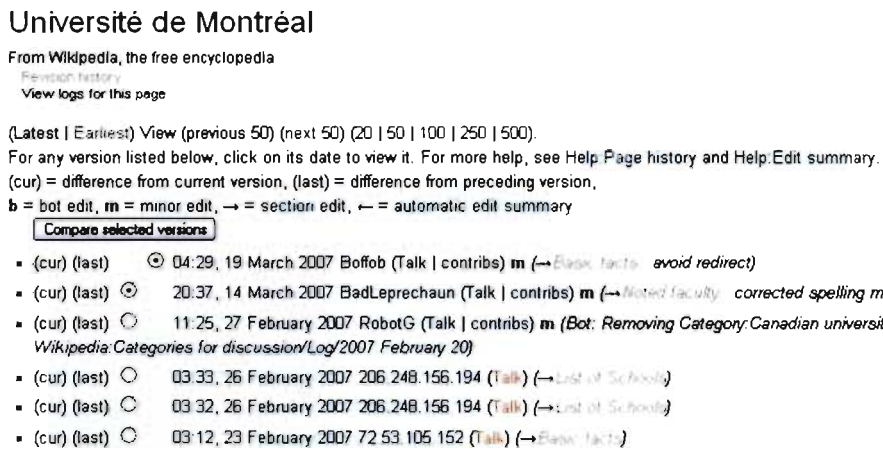

Figure 1.2 – Historique des modifications (*format HTML*)

– liens internes (ou hyperliens) permettant d'accéder aux autres pages principales.

Nous nous sommes limités aux caractéristiques communes à plusieurs engins Wikis, d'autres options existent suivant le moteur de Wiki utilisé. On peut citer notamment la page de discussion qu'on retrouve sur les Wikis fonctionnant avec le moteur de Wiki "MediaWiki" comme l'encyclopédie Wikipedia. Ce type de page est reliée à la page principale et comprend toutes les discussions sur le contenu de la page principale correspondante. La page de discussion possède également son propre historique des modifications.

1.2 Problématique abordée au cours du mémoire

Dans ce mémoire, nous nous sommes intéressée aux Wikis de façon dynamique, plus précisément nous nous sommes préoccupés de la façon dont la qualité est assurée et maintenue dans un document résultant d'une collaboration entre internautes.

En effet, comme nous le verrons dans le prochain chapitre, la majorité des études qui se sont focalisées sur les Wikis ont ignoré leur aspect dynamique : comment on est passé de la première version à la version courante du document ? Or les Wikis (et spécifiquement Wikipedia) offrent des outils d'édition collaborative sans précédent [10] tout en imposant des contraintes inconnues auparavant. L'interaction de ces deux facteurs résulte en l'émergence de mécanismes de collaboration nouveaux comme la recherche d'un consensus et le maintien de la cohérence.

Ainsi, il est fort intéressant de savoir comment un consensus émerge et comment la cohérence est assurée sur le contenu d'un article de Wikipedia quand on sait que les contributeurs ne se connaissent pas en général et que ceci empêche l'émergence d'une autorité pouvant servir de référence.

Une façon de tracer l'évolution du contenu d'un article est de se poser la question en termes quantitatifs. Autrement dit, qui parmi les auteurs est à l'origine d'un passage du texte final ? Où sont les autres passages du même auteur ? Quel est leur "poids" dans le texte final ? Qu'est-ce que sont devenus les autres contributions du même auteur qui ne sont pas arrivées dans la version finale ? Quel est l'âge (en nombre de versions) des informations incluses dans la version finale ?

Cet angle de vue a été adopté par certains chercheurs comme nous le verrons dans le prochain chapitre, notamment les concepteurs du système *History Flow* qui permet de visualiser les contributions de chaque utilisateur, de suivre tous les changements effectués sur un texte dès son introduction, etc. Un des scénarios (ou mécanismes

de collaboration) facile à détecter à l'aide de *History Flow* est le *Vandalisme* qui consiste à des modifications non autorisées et non désirées du contenu des pages et qui se manifeste sous différentes formes, entre autre par la suppression d'une partie ou totale du texte d'un document. La figure 1.3 à la page 6 illustre ce scénario (cadre jaune : suppression totale du contenu et la restauration ou retour à la version précédente).

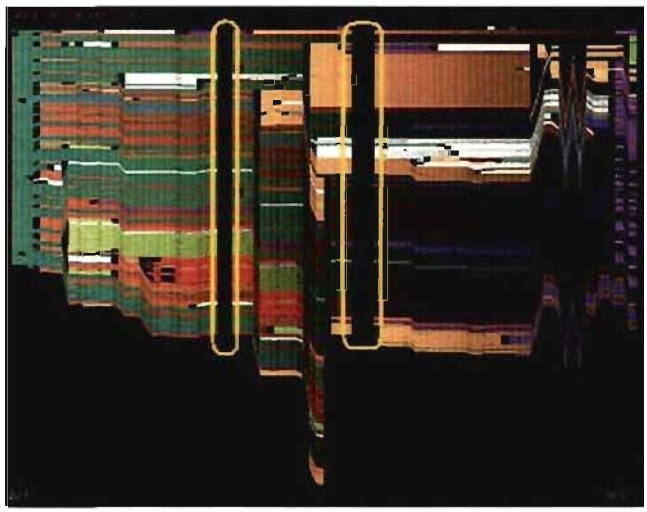

Figure 1.3 – Vandalisme - Restauration, History flow

Pour notre part, nous avons adopté une vue complémentaire à celle de cet outil. Elle s'intéresse d'avantage aux aspects qualitatifs de l'évolution. Autrement dit, quel est le rôle d'une intervention donnée d'un auteur dans l'élaboration de l'idée sous-jacente dans le texte? Dans la poursuite de cette interrogation, nous nous sommes concentrés sur les aspects sémantiques dans les rapports entre interventions individuelles ou, plus spécifiquement, entre une intervention et le texte qui la précède.

Notre hypothèse est qu'en observant les séquences d'interventions d'un nombre suffisant d'articles (en réalité d'idées ou paragraphes) et en ne retenant que les plus fréquentes, nous serons en mesure de détecter les manifestations des mécanismes de

collaboration visés. Pour ce faire, le principe linguistique de cohérence textuelle sera notre guide car nous supposons qu'à tout instant, les auteurs cherchent à produire un texte cohérent et par là, mettent explicitement les liens existant entre le contenu de leur intervention et le texte entourant.

1.3 Cadre de ce travail

Ce travail a été effectué dans le cadre du projet "Étude de l'émergence de nouveaux principes de gouvernance au sein des collaborations appuyées par les TIC" mené conjointement par des chercheurs du Département de communication et du Département d'informatique et de recherche opérationnelle de l'Université de Montréal. Le projet visait à identifier les *principes de gouvernance* qui régissent la collaboration au sein des nouvelles *communautés en ligne*[4]. Nous entendons par principes de gouvernance, les règles (implicites et explicites) d'intervention ainsi que les pratiques de discipline et de régulation des interactions. Nous nous sommes intéressés particulièrement aux principes de gouvernance qui se constituent autour des sites Wikis.

En effet, ayant constaté que lors de la rédaction du document les wikistes appliquent certaines règles afin de produire un texte cohérent, les chercheurs ont voulu identifier ces règles (ou plus généralement ces mécanismes de collaboration) dans les pages Wiki. Pour ça les chercheurs avaient besoin d'un outil performant, automatique et prenant en compte plusieurs versions d'un article (la fonctionnalité "diff" des sites Wiki et spécialement de Wikipedia, ne permettant que la comparaison de deux versions et l'outil **history flow** pour visualiser plusieurs versions, voire même l'historique des modifications en entier, présentant d'autres limites comme le manque de l'aspect sémantique des interventions).

[4] Ensemble de personnes reliées par ordinateur dans le cyberespace, qui se rencontrent et ont des échanges par l'intermédiaire d'un réseau informatique, tel l'Internet, et qui partagent un intérêt commun. [6]

Ainsi nous avons proposé aux spécialistes en communication, un outil personnalisé de visualisation des articles de Wikipédia : *historyDiff* ainsi qu'une méthodologie automatique d'identification des règles recherchées qui prend en compte un ou plusieurs historique(s) des modifications, intègre l'aspect sémantique des interventions et qui consiste en une extraction des régularités (ou séquences de catégories d'interventions répétitives), manifestations des règles visés.

De plus l'outil *historyMiner*, qui est une mise en pratique de cette approche, a été implémenté permettant ainsi aux chercheurs en communication d'extraire et de visualiser les motifs de collaboration qu'ils pouvaient alors interpréter afin de formuler des règles d'intervention (ou des mécanismes de collaboration) propres aux sites Wikis.

1.4 Méthodologie

Les mécanismes de collaboration (comme *Vandalisme* $\Rightarrow$ *Restauration*) se manifestent par des séquences d'interventions des utilisateurs typiques (exemple : *une suppression totale du contenu d'un article* suivi d'*un retour à la version précédente* ou *une insertion de mots vulgaires et insultes* suivi d'*une suppression du même texte*) [21]. Afin de permettre aux chercheurs en communication d'identifier ces mécanismes, nous avons extrait des séquences d'interventions typiques, manifestations des règles recherchés et présenté ces dernières aux spécialistes en communication afin qu'ils puissent les interpréter.

Plus généralement, nous nous inspirons des techniques de la *fouille de données* [9]. En effet, comme nous allons le voir dans le chapitre qui suit, la fouille de données consiste en la découverte de nouvelles connaissances comme des modèles ou des règles, dans une grande variétés de données à analyser, entre autres des bases de données relationnelles, des bases de données transactionnelles ou encore des bases de données textuelles comme les pages html/xml. En outre, la technique de *règles*

d'association de ce processus pour rechercher des règles se fait en deux étapes : tout d'abord des motifs sont générés et puis les règles correspondantes sont formulées. Une variation de cette approche : les *motifs séquentiels* [1] permet en plus de prendre en compte l'aspect temporel entre les données à analyser.

Dans notre cas, nous devions analyser des données textuelles, les interventions des participants faisant partie des historiques des modifications[5] et nous recherchions des séquences d'interventions répétitives qui réalisent des règles régissant la collaboration sur un site Wiki.

Plus spécifiquement, notre travail a porté sur trois points : Tout d'abord, nous nous sommes focalisés sur l'évolution d'un échantillon représentatif de pages d'un site Wiki afin d'identifier les divers types d'interventions des participants. À cette fin, le logiciel "*historyDiff*" (voir la figure 1.4 à la page 10) permettant de visualiser le texte de plusieurs versions d'un historique des modifications avec une mise en évidence des interventions a été implémenté. Les couleurs vert, rouge et jaune correspondant respectivement à l'insertion, suppression et déplacement d'un bloc de texte.

Avec ce logiciel, nous pouvions facilement identifier les deux types d'opérations de modification principaux à savoir : l'insertion et la suppression. Cependant ces données ne nous donnaient pas beaucoup d'informations quand aux mécanismes de collaboration appliqués sur les Wikis. Un autre constat de cette visualisation de l'historique avec *historyDiff* est que les auteurs cherchent à produire un texte cohérent en procédant notamment par jonction (ou utilisation des connecteurs) afin d'indiquer explicitement les relations entre des phrases ou paragraphes de leur intervention et ceux existant auparavant dans le document.

[5]Dans la section introductive sur les sites Wikis, nous avons vu que les interventions des participants pouvaient être extraites en calculant la différence entre chaque paire de versions successives du document ou page d'un Wiki.

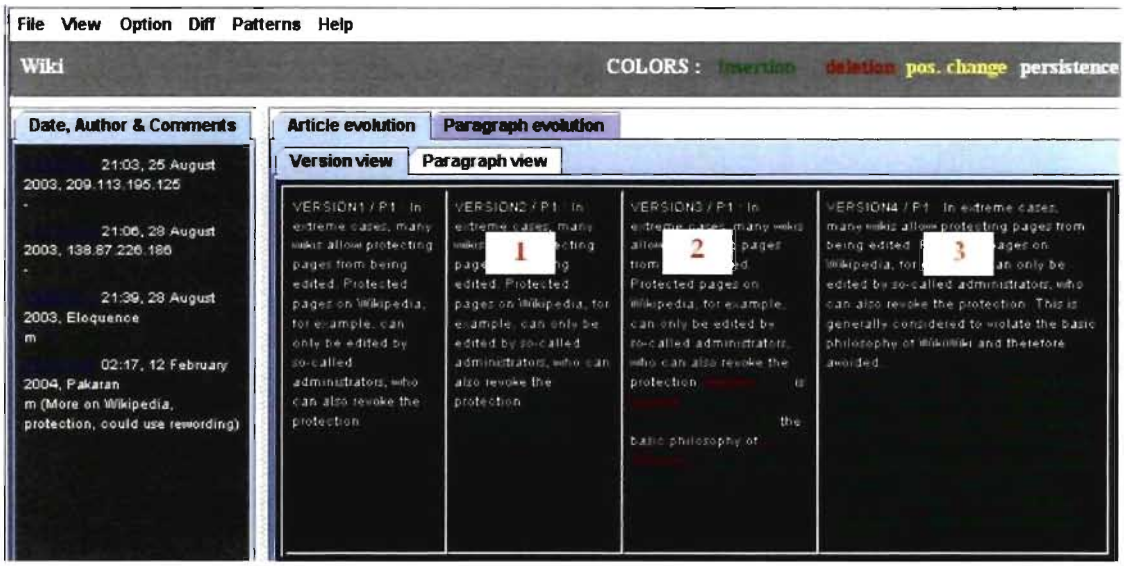

Figure 1.4 – Aperçu de l'article "Wiki" dans *historyDiff*

Pour la suite, nous avons effectué des recherches sur la notion de cohérence textuelle dont un compte-rendu est dans le prochain chapitre. Puis, en considérant les caractéristiques des interventions sur un site Wiki (les différents types d'opérations de modification et les procédés appliqués par les utilisateurs pour garder le texte cohérent) et en s'inspirant des recherches sur la cohérence textuelle, nous avons élaboré une taxonomie des interventions.

La taxonomie ainsi formée durant la première étape a par la suite été utilisée pour classer, de façon automatique, des structures d'interventions tirées des historiques des modifications.

Pour ce faire, nous avons premièrement considéré différentes versions d'un corpus de pages de Wikipedia, identifié les interventions de chaque participant puis constitué des séquences d'interventions qui reflètent aussi bien les contenus des interventions (les opérations de modification et les textes des contributions) que les rapports entre les interventions et le texte précédent (relations sémantiques exprimées par

des marqueurs). Les interventions des séquences résultantes ont été par la suite catégorisées en utilisant la taxonomie d'interventions.

Considérons les 3 interventions mises en évidence sur la figure 1.4 à la page 10 et plus précisément les opérations de modification (ajout ou suppression) ainsi que les connecteurs utilisés, nous pouvons constater que les interventions 1 et 3 ont consisté en un ajout d'un texte mais de taille différente. De plus les deux interventions contiennent un marqueur de restriction ("But" pour l'intervention 1 et "However" pour l'intervention 3). En se basant sur ces renseignements, nous pouvons classer l'intervention 1 dans la catégorie d'intervention : "Restriction légère" et l'intervention 3 dans la catégorie : "Restriction argumentée".

Le résultat de la deuxième phase (catégorisation des interventions) a enfin constitué l'entrée d'un processus d'analyse à l'aide d'une technique de fouille de données afin d'en extraire les manifestations des règles visés, c'est-à-dire, des configurations locales de diverses catégories d'interventions qui apparaissent fréquemment dans les structures globales. Étant donné la structure séquentielle des données initiales, il a fallu appliquer l'approche de motifs séquentiels, plus précisément nous avons utilisé l'algorithme Apriori dont les principes se trouvent détaillés et illustrés dans le deuxième chapitre de ce rapport.

Une fois les motifs extraits, ils sont présentés à l'utilisateur, en l'occurrence les chercheurs en communication afin qu'il puisse les interpréter et formuler les règles ou mécanismes de collaboration propres aux Wikis.

1.5 Organisation du mémoire

La suite de ce mémoire est organisé de la manière suivante.

Dans le chapitre 2, nous allons présenter la problématique abordée dans ce mémoire. Nous allons donner différentes définitions afin de poser la problématique de la recherche des motifs de collaboration. Par la suite, nous allons examiner les différents travaux existants autour de cette problématique. Nous présentons tout d'abord les travaux qui se sont focalisés sur la qualité des contenus sur les Wikis. Par la suite, nous voyons un aperçu de la notion de cohérence textuelle et des règles pour l'assurer. Et enfin nous faisons un rappel sur les travaux autour des règles d'association ainsi que sur les motifs séquentiels.

Le chapitre 3 présente notre approche pour traiter la problématique de la recherche des motifs de collaboration fréquents. Nous donnons son architecture et expliquons son fonctionnement.

Le chapitre 4 est consacré à *historyMiner*, l'outil mettant en pratique la méthodologie proposée. Nous allons parler de son implémentation et allons expliquer son fonctionnement.

Le chapitre 5 montre les expérimentations effectuées pour la validation de notre méthodologie. À cette fin, l'outil *historyMiner* a été utilisé pour extraire des motifs de collaboration fréquents dans un ensemble d'articles de l'encyclopédie Wikipedia (version anglaise). Nous présentons les deux corpus étudiés ainsi que les résultats obtenus.

Enfin, le chapitre 6 résume le travail effectué et donne quelques pistes à explorer dans le futur.

CHAPITRE 2

PROBLÉMATIQUE ET TRAVAUX ANTÉRIEURS

2.1 Présentation de la problématique étudiée

Dans ce mémoire, nous allons aborder la problématique de la recherche des mécanismes de collaboration au sein des sites Web Wikis. Nous entendons par mécanismes de collaboration, les comportements typiques des wikistes lors de la rédaction des contenus comme la façon dont ils procèdent pour assurer la qualité des contenus : comment obtiennent-ils un consensus sur des contenus controversés ou encore comment assurent-ils la cohérence du texte ?

Or, ces mécanismes de collaboration se manifestent généralement par des séquences d'interventions des utilisateurs que l'on peut identifier dans les historiques des modifications [21]. La problématique de la recherche des scénarios collaboratifs peut ainsi être scindée en deux partie : tout d'abord l'identification des séquences d'interventions typiques dans les logs des pages et par la suite l'interprétation de ces motifs par un expert afin de déterminer les mécanismes recherchés.

De plus, les interventions individuelles sont spécifiques au texte (de la page). Afin d'identifier des mécanismes de collaboration représentatifs, nous allons considérer les catégories d'interventions qui sont, contrairement aux interventions, plus générales.

Le présent chapitre est consacré au problème de la recherche de séquences de catégories d'interventions répétitives dans les historiques des modifications des pages des sites Wikis ou recherche de motifs de collaboration fréquents. Après une première partie consacrée aux définitions nous allons poser la problématique de recherche de motifs de collaboration fréquents.

2.1.1 Définitions

Dans cette partie, nous allons présenter différentes définitions pour le processus de recherche de motifs de collaboration fréquents. Après une première partie introductive sur les différents types d'opérations de modification, la notion d'intervention et celle de la séquence d'interventions, nous présenterons le concept de catégorie d'interventions et de la séquence de catégories d'interventions. Nous verrons par la suite les relations de spécialisation et de couverture pour enfin définir la notion de motif de collaboration (fréquent).

2.1.1.1 Intervention

Comme nous l'avons souligné dans notre introduction, chaque page d'un site Wiki a un historique des modifications qui comprend différentes informations sur les transformations subies par cette page depuis la première version jusqu'à une version finale (en l'occurrence la version en cours). Parmi ces informations, on retrouve notamment le texte intégral de chaque version de l'article. Les textes des différentes versions d'une page Wiki peuvent être considérés comme des fichiers textes. Ainsi, la mise en évidence des opérations effectuées par un contributeur pour modifier le texte d'une version revient à calculer la différence entre le texte de cette version et celui de la version suivante.

Toutes les modifications appliquées à un fichier texte peuvent être décrites par deux types d'opération de modification suivants :

- $insert(t_x)$: insertion du texte t_x
- $delete(t_x)$: suppression du texte t_x

À partir de ces deux opérations de modification, un *script de modification* est définit comme étant une séquence d'opérations de modifications qui transforme un fichier texte *f1* en un fichier texte *f2*.

En considérant les opérations effectuées par les utilisateurs des sites Wikis pour modifier une page, nous pouvons définir une *intervention* comme étant une sé-

quence d'opérations de modification effectuées par le wikiste pour modifier une version d'une page.

De plus, le texte d'une page sur un Wiki est structuré en différents niveaux, à savoir : mot, groupe de mots, phrase, paragraphe. Or, le dernier élément (paragraphe) est considéré en linguistique comme étant l'unité de sens homogène autrement-dit centré sur une seule idée. Ainsi l'ajout d'un nouveau paragraphe dans une page d'un site Wiki peut être traduit comme l'ajout d'une nouvelle idée, le découpage d'un paragraphe en deux paragraphes, comme la division d'une idée composée en deux idées plus simples, l'ajout d'une phrase dans un paragraphe comme une illustration d'une idée par des faits, etc.

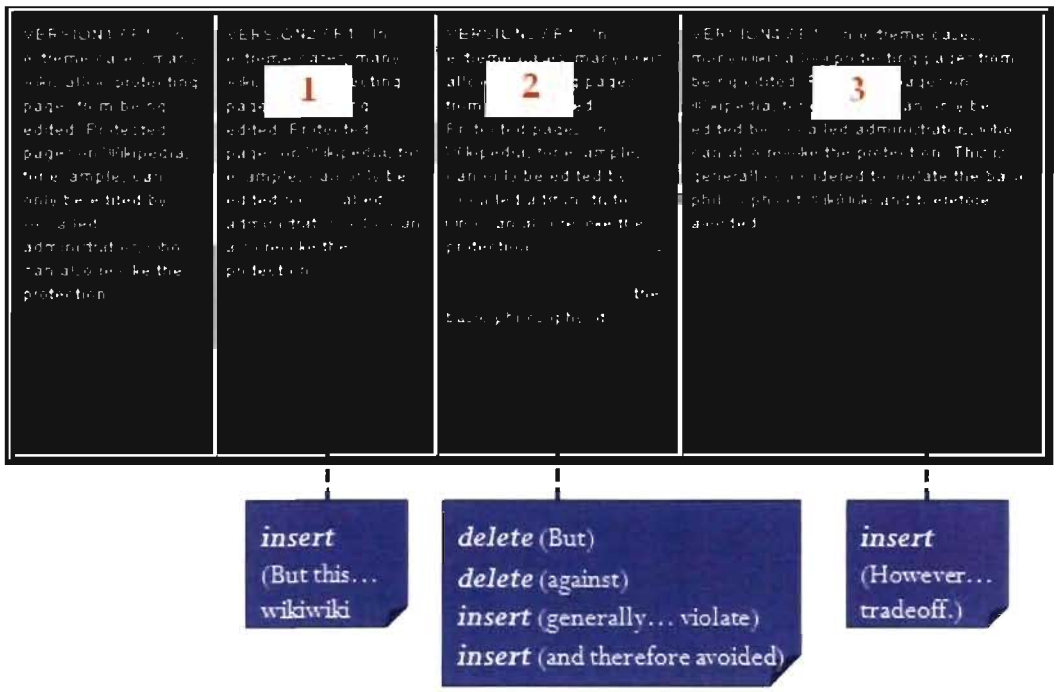

Figure 2.1 – Identification d'une intervention : différences entre versions

Dans la figure 2.1 à la page 15 qui illustre une partie de l'historique des modifications d'un paragraphe (de la page "Wiki" de la version anglaise de Wikipedia), nous pouvons identifier trois interventions obtenues par le calcul de la différence

entre les 4 versions successives du paragraphe. Nous pouvons également constater que la deuxième intervention, contrairement aux deux autres, est une séquence de plusieurs opérations de modification.

Considérons deux versions successives d'une page d'un site Wiki, formellement, l'intervention d'un utilisateur pour modifier un paragraphe de l'ancienne version en un paragraphe de la nouvelle version est définie comme suit :

Définition 1 : Intervention (ι)

Soient T_i et T_{i+1} les textes respectifs de V_i et V_{i+1}, deux versions successives d'une page d'un site Wiki tel que i est l'estampille (T_i ou T_{i+1} peut être vide). Soient t_j^i le texte de P_j^i (le $j^{ième}$ paragraphe de T_i), $t_j^i \leq T_i$ (t_j^i compose partiellement ou intégralement T_i) et t_k^{i+1} le texte de P_k^{i+1} (le $k^{ième}$ paragraphe de T_{i+1}), $t_k^{i+1} \leq T_{i+1}$ tel que P_k^{i+1} est le correspondant de P_j^i (les deux paragraphes sont similaires à quelques mots près, voir le chapitre 4). Une intervention subie par le paragraphe P_j^i est notée ι, c'est une séquence, non vide, d'opérations de modification $< o_1, o_2, \ldots, o_m >$ effectuées par un utilisateur pour transformer t_j^i en t_k^{i+1}.

$\iota = diff(t_k^{i+1}, t_j^i) =< o_1, o_2, \ldots, o_m >$

où o_l $(l = 1 \ldots m)$ est une opération de modification.

L'ajout (ou suppression) d'un paragraphe est un cas particulier. Dans ce cas précis, l'intervention sera une séquence formée d'une seule opération de modification : $\iota = insert(t_x)$ ou $\iota = delete(t_x)$ avec t_x, le texte du paragraphe en question. De plus, si au lieu d'avoir deux versions d'un texte, on en a plusieurs, ce procédé d'identification d'une intervention entre deux versions successives d'un paragraphe (d'une page Wiki) peut être répété sur chaque paire de versions successives de ce paragraphe. Ceci correspond à l'évolution, dans le temps, d'un paragraphe (d'une page d'un site Wiki) et consiste par conséquent en une liste ou séquence d'interventions.

Définition 2 : Séquence d'interventions (I)

Soit $< T_1, T_2, T_3, \ldots, T_{n-1}, T_n >$ la séquence des textes de n versions d'une page d'un site Wiki. Soit $< t^1_{j_1}, t^2_{j_2}, t^3_{j_3}, \ldots, t^{n-1}_{j_{n-1}}, t^n_{j_n} >$ la séquence des textes de n paragraphes tel que :

- $t^i_{j_i}$ est le texte de $P^i_{j_i}$ (le $j^{ième}$ paragraphe de T_i) avec $1 \leq i \leq n$
- $P^{i+1}_{j_{i+1}}$ est le correspondant de $P^i_{j_i}$ avec $1 \leq i < n$

Une séquence d'interventions est notée I, c'est une liste ordonnée, non vide, d'interventions $<\iota_1, \iota_2, \ldots, \iota_{n-1}>$ opérées par différents utilisateurs pour transformer un paragraphe d'une page d'un site Wiki depuis une version d'*origine* (la version qui précède la première version dans laquelle ce paragraphe a été introduit) jusqu'à une version *finale* (la version suivant la dernière version dans laquelle ce paragraphe apparaît).

$I =< diff(t^2_{j_2}, t^1_{j_1}), diff(t^3_{j_3}, t^2_{j_2}), \ldots, diff(t^n_{j_n}, t^{n-1}_{j_{n-1}}) >=< \iota_1, \iota_2, \ldots, \iota_{n-1} >$

où ι_k $(k = 1 \ldots n-1)$ est une intervention.

2.1.1.2 Catégorie d'interventions

Les opérations de modification composant les interventions des utilisateurs d'un site Wiki peuvent être considérées sous plusieurs angles dont deux nous intéressent particulièrement ici. Nous les avons appelés de façon approximative, la *"syntaxique"* et la *"sémantique"*.

En effet, chaque opération de modification composant une intervention peut être classée dans une catégorie suivant deux descriptions : le type d'opération de modification (ou aspect syntaxique) ainsi que la relation sémantique (avec l'intervention précédente) exprimée par un connecteur (ou aspect sémantique). Ainsi une intervention correspond à une ou plusieurs catégories auxquelles appartiennent les opérations de modification qui la composent. La figure ci-dessous illustre une catégorisation possible des interventions identifiées dans la figure 2.2 à la page 18.

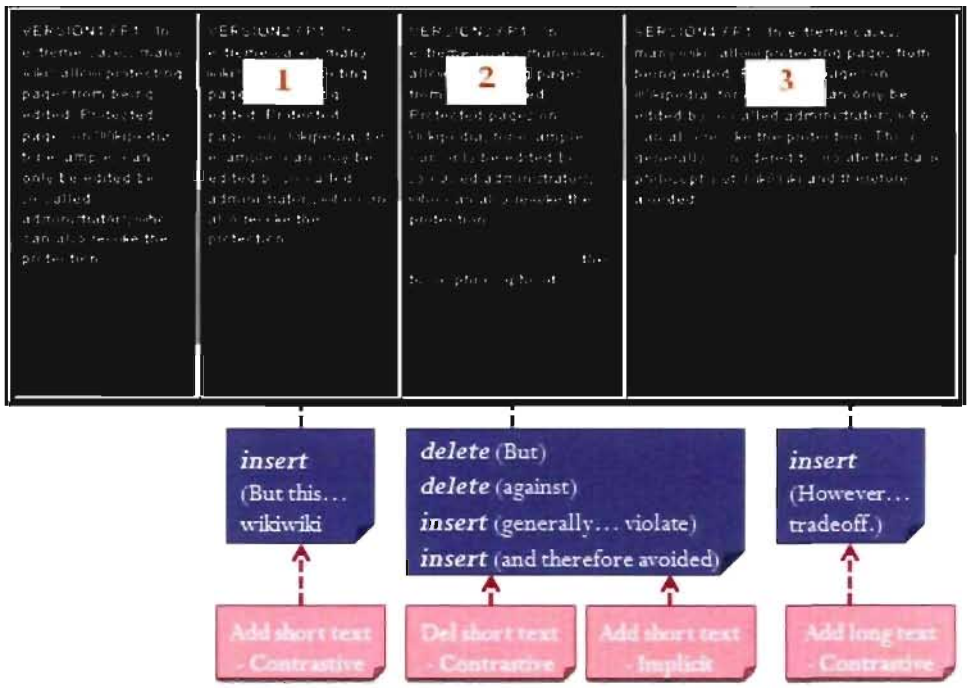

Figure 2.2 – Description générale d'une intervention

Définition 3 : Catégorie d'interventions (c)

Soient A_{sy} et A_{se}, deux angles *syntaxique* et *sémantique*, pour décrire les opérations de modification composant les interventions des contributeurs lors de la modification du contenu d'une page d'un site Wiki. Une catégorie d'interventions est notée c, c'est un couple (x,y) tel que x est une valeur de l'angle syntaxe et y est une valeur de l'angle sémantique.

$c = (x,y) : x \in A_{sy}$ et $y \in A_{se}$

De plus, les deux valeurs de l'angle syntaxique : ajout d'un texte et suppression d'un texte peuvent être scindées en deux nouvelles valeurs chacune par considération de la taille du texte. L'ensemble de toutes ces valeurs forme une classification (ou taxonomie) à deux niveaux dont les valeurs sont reliées par une relation "*is-a*" ou "*est une sorte de*". Ainsi "Ajout d'un long texte" est une sorte de "Ajout d'un texte".

Les valeurs de l'angle sémantique s'inspirent des relations sémantiques pouvant exister entre les différentes parties d'un texte (comme la cause, l'addition, etc.). Celle-ci est également une classification dont les valeurs sont reliées par une relation "*is-a*" ou "*est une sorte de*". Dans la troisième section de ce chapitre nous verrons le développement détaillé de ces deux classifications.

Supposons deux taxonomies (ou classifications), la syntaxique et la sémantique, pour décrire les opérations de modification composant les interventions des contributeurs lors de la modification du contenu d'une page d'un site Wiki. Formellement, l'espace de toutes les valeurs couvert par ces deux taxonomies forme une *taxonomie des interventions* et est défini comme suit :

Définition 4 : Taxonomie des interventions (τ)
Soient T_{sy} (la taxonomie "*syntaxique*") et T_{se} (la taxonomie "*sémantique*"), deux classifications des opérations de modification. Une taxonomie des interventions sur un Wiki est notée τ, c'est un ensemble non vide de couples $\{(x_1,y_1)\dots(x_n,y_m)\}$ où x_i $(1 \leq i \leq n)$ est une valeur de la taxonomie syntaxique et y_j $(1 \leq j \leq m)$ est une valeur de la taxonomie sémantique, qui sont hiérarchisés (ou à différents niveaux) et sont reliés par une relation "i*s-a*" ou "*est une sorte de*".

Considérant cette taxonomie des interventions sur un Wiki, une catégorie d'interventions est donc définie comme étant un élément de cette taxonomie. De plus la *profondeur d'une catégorie d'interventions* est définie comme suit :

Définition 5 : Profondeur d'une catégorie d'interventions ($d(c)$)
Soient τ, une taxonomie des interventions et c, une catégorie d'interventions $(c \in \tau)$ telles que définies précédemment. La profondeur de c est notée $d(c)$, c'est le **niveau** auquel est situé c dans τ. Cette valeur vaut "0" si c'est la racine.

Définition 6 : Relation *is-a* dans la taxonomie des interventions ($\leq_\tau$)

Soit τ, une taxonomie des interventions telle que définie précédemment. Soient c_1 et c_2, deux éléments faisant partie de cette taxonomie (autrement-dit c_1 et c_2 sont deux catégories d'interventions). Une relation is-a entre c_1 et c_2 est notée $\leq_\tau$, elle stipule que c_1 est une catégorie d'interventions plus spécifique de c_2.

Définition 7 : Categorieset (s)

Un *categorieset*, noté s, est un ensemble non vide $\{c_1, c_2, \ldots, c_m\}$ où c_i $(i = 1 \ldots m)$ est une catégorie d'interventions.

Définition 8 : Séquence de categoriesets (C)

Une séquence de categoriesets est notée C, c'est une liste ordonnée non vide $<s_1, s_2, \ldots, s_k>$ où s_i $(i = 1 \ldots k)$ est un *categorieset*.

$$C =< \{c_1^1, c_2^1, \ldots, c_{p_1}^1\}, \{c_1^2, \ldots, c_{p_2}^2\}, \ldots, \{c_1^k, \ldots, c_{p_k}^k\} >=< s_1, s_2, \ldots, s_k >$$

Définition 9 : Rang d'une séquence de categoriesets ($rang(C)$)

Soit C, une séquence de categoriesets telle que définie ci-dessus. Le rang de C est notée $rang(C)$, c'est la **somme** des profondeurs de toutes les catégories d'interventions comprises dans la séquence C :

$rang(C) = \sum_{i=1}^{k} \sum_{j=1}^{p_i} d(c_j^i)$

où $d(c_j^i)$ est la profondeur de la catégorie d'interventions c_j^i

Définition 10 : Relation de spécialisation ($\ll$)

Soient M et S deux séquences telles que définies ci-dessous :

$$M =< \overbrace{\{b_1^1, b_2^1, \ldots, b_{q_1}^1\}}^{m_1}, \overbrace{\{b_1^2, \ldots, b_{q_2}^2\}}^{m_2}, \ldots, \overbrace{\{b_1^l, \ldots, b_{q_l}^l\}}^{m_l} >$$

$$S =< \underbrace{\{c_1^1, c_2^1, \ldots, c_{p_1}^1\}}_{s_1}, \underbrace{\{\ldots\}}_{s_2}, \underbrace{\{c_1^3, \ldots, c_{p_3}^3\}}_{s_3}, \ldots, \underbrace{\{c_1^k, \ldots, c_{p_k}^k\}}_{s_k} >$$

(avec $s_1 \uparrow\ll$, $s_2 \nwarrow\ll$, $s_3 \nwarrow\ll$)

Soient m_i, un *categorieset* appartenant à M et s_j, un *categorieset* appartenant à S. Soit $\leq_\tau$, la relation *is-a* dans la taxonomie des interventions. Formellement, une relation de spécialisation entre les *categoriesets* m_i et s_j composant ces deux séquences est notée $\ll$, elle est définie comme suit : pour chaque catégorie d'interventions dans m_i, il existe une catégorie d'interventions dans s_j qui la spécialise (relation *is-a* de la taxonomie des interventions).

s_j spécialise m_i ($s_j \ll m_i$) ssi $\forall x \in [1 \ldots q_i]\ \exists y \in [1 \ldots p_j] : c_y^j \leq_\tau b_x^i$
avec $\leq_\tau$, la relation *is-a* dans la taxonomie des interventions τ.

À partir de cette relation de spécialisation, une relation de couverture entre les deux séquences de catégoriesets M et S est définit comme suit :

Définition 11 : Relation de couverture ($\sqsubseteq$)
Soient M et S deux séquences de catégoriesets telles que nous les avons définies ci-dessus.
M couvre S ($M \sqsubseteq S$) ssi :

1. Correspondance : $\forall i \in [1, |M|]\ \exists j \in [1, |S|]$ tel que $s_j \ll m_i$ (s_j specialises m_i).
2. Ordre : soient $j_1, j_2, \ldots, j_l$ des entiers tel que $s_{j_i} \ll m_i$ (avec i de 1 à l) (correspondance) $\Rightarrow j_1 < j_2 < \ldots < j_l$.

2.1.1.3 Motif de collaboration

Afin d'identifier l'évolution d'un paragraphe (d'une page d'un site Wiki), un identifiant unique ***PID*** lui est associé. Cet identifiant est donc associé à la séquence d'interventions correspondant à ce paragraphe et par conséquent à une séquence spécifique de categoriesets.

Définition 12 : Séquence de données (S)
Soit $I_{PID} = \langle \iota_1, \iota_2, ..., \iota_k \rangle$ une séquence d'interventions correspondant à l'évolution d'un paragraphe spécifique. Une séquence de données est notée S, c'est une sé-

quence de categoriesets, non vide $\langle s_1, s_2, \ldots, s_k \rangle$ correspondant à la séquence d'interventions I_{PID} (autrement-dit, c'est la séquence d'intervention I_{PID} dans laquelle on a remplacé chaque intervention ι_i par s_i, l'ensemble des différentes catégories d'interventions auxquelles appartiennent les opérations de modification composant l'intervention ι_i). C'est un cas particulier d'une séquence de categoriesets car il correspond aux interventions d'un paragraphe bien déterminé.

$S = (PID, \langle s_1, s_2, \ldots, s_k \rangle)$
où PID est l'identifiant unique du paragraphe et s_i $(i = 1 \ldots k)$ est un categorieset.

Définition 13 : Motif de collaboration (M)
Un motif de collaboration est notée M, il s'agit de toute séquence de catégoriesets $\langle m_1, m_2, \ldots, m_l \rangle$ pouvant être générée à partir d'un ensemble de catégories d'interventions (taxonomie des interventions).

$M = \langle m_i \rangle_{i=1\ldots l}$ où $m_i = \left\{ b_j^i \right\}_{j=1\ldots qi}$ et b_j^i est une *catégorie d'interventions.*

Un paragraphe (d'une page d'un site Wiki) *supporte* un motif de collaboration M (ou un paragraphe fait partie du support de M) si M couvre la *séquence de données* S correspondant à ce paragraphe (voir la relation de couverture définie plus haut). Formellement, le *support* d'un motif de collaboration peut être défini comme suit :

Définition 14 : Support d'un motif de collaboration ($supp(M)$)
Soit M, un motif de collaboration. Le support de M est notée $supp(M)$, il est calculé comme étant le pourcentage des paragraphes (d'une page ou plus généralement de plusieurs pages d'un site Wiki) dont les séquences de données sont couverts par M.

2.1.2 Problème de recherche de motifs de collaboration fréquents

En considérant les différentes définitions vues plus haut, la problématique abordée dans ce mémoire peut être résumée ainsi : soient H, un ensemble d'*historiques des modifications* et ***minSupp*** le support minimum fixé par l'utilisateur. L'ensemble

$M^{H,minSupp}$ des *motifs de collaborations fréquents* (ou *séquences de catégoriesets fréquents*) est l'ensemble de tous les motifs de collaboration $\{mc_1, mc_2, \ldots, mc_n\}$ tel que pour chaque motif de collaboration mc_i, avec i de 1 à n, dans $M^{H,minSupp}$, mc_i vérifie (ou a un support supérieur à) *minSupp*.

Le problème de la recherche de motifs dans la rédaction collaborative sur les sites Wiki (titre de ce mémoire) consiste à trouver l'ensemble $M^{H,minSupp}$ (autrement-dit, tous les motifs dont le support est supérieur à *minSupp*).

Dans la section qui suit, nous allons examiner les travaux reliés à la recherche des motifs de collaboration fréquents dans les historiques des modifications des pages d'un site Web Wiki.

2.2 Aperçu des travaux antérieurs

Depuis quelques années, les sites Wikis et particulièrement l'encyclopédie Wikipedia ont suscité beaucoup d'intérêts dans la communauté scientifique. En effet, deux conférences internationales WikiSym[1] et Wikimania[2] qui lui sont consacrées sont à leur troisième édition et plusieurs études [18, 22, 23] sont réalisées chaque année permettant de mieux comprendre cette récente technologie.

Selon Wikimedia [23], les travaux sur les Wikis peuvent être classés en 19 catégories dont la "Qualité", qui regroupe tous les travaux dans lesquels les auteurs, tout comme nous, se sont intéressés à la qualité de l'information sur un site Wiki. Dans ce mémoire, nous complétons ces études en nous intéressant particulièrement au rôle d'une intervention donnée d'un auteur dans le développement des contenus sur un site Wiki (notamment l'encyclopédie Wikipedia) et en dégageant ainsi de nouveaux règles ou mécanismes de collaborations régissant une telle rédaction.

[1] WikiSym - http ://www.wikisym.org/
[2] Wikimania - http ://wikimania2007.wikimedia.org/

Pour cela, nous allons nous inspirer du principe linguistique de la cohérence textuelle ainsi que des différents procédés utilisés pour l'assurer afin de déterminer l'aspect sémantique d'une intervention autrement-dit son rôle exact dans l'élaboration de l'idée sous-jacente dans le texte.

En se basant entre autres, sur cet aspect sémantique, nous allons catégoriser différentes interventions d'auteurs, identifiées auparavant dans les historiques des modifications des pages d'un Wiki et organisées dans des séquences. Puis nous allons y extraire des motifs de collaborations fréquents qui sont des ébauches des règles régissant la rédaction collaborative des contenus sur ces outils. À cette fin, l'approche Apriori, une technique de la fouille de données pour extraire des motifs fréquents, sera utilisé.

La présente section est organisée de la manière suivante. Dans la première partie nous allons présenter des travaux reliés à la recherche des mécanismes de collaboration dans les sites Wikis ou plus généralement à la qualité des contenus sur les Wikis. Nous introduirons au cours de la deuxième partie la notion de cohérence textuelle et les procédés pour l'assurer. Dans la troisième partie nous faisons un rappel sur les travaux de la fouille de données.

2.2.1 Qualité des informations sur un site Wiki

La recherche des mécanismes de collaboration s'inscrit dans la catégorie des travaux qui visent à analyser la qualité des informations sur un site Wiki. Dans cette section nous allons parler des différents travaux antérieurs de cette catégorie. Pour chaque travail, nous verrons entre autre les techniques utilisées, les résultats obtenus par les chercheurs et enfin les limitations que nous avons relevées.

2.2.1.1 Évaluation de la qualité d'information par visualisation des flux des modifications

Comme nous l'avons noté dans notre introduction, la qualité de l'information sur les Wikis et plus particulièrement celle des contenus de l'encyclopédie Wikipedia est un point très discuté dans la communauté. En effet, plusieurs travaux qui s'attachent à mesurer un mouvement qualitatif (de façon manuelle ou automatique) des articles de Wikipedia ont été effectués. *F. Viégas et al.* [21] ont proposé *history flow*, un outil de visualisation des flux des modifications, en l'occurrence le journal (ou historique des modifications) d'une page sur un site Wiki.

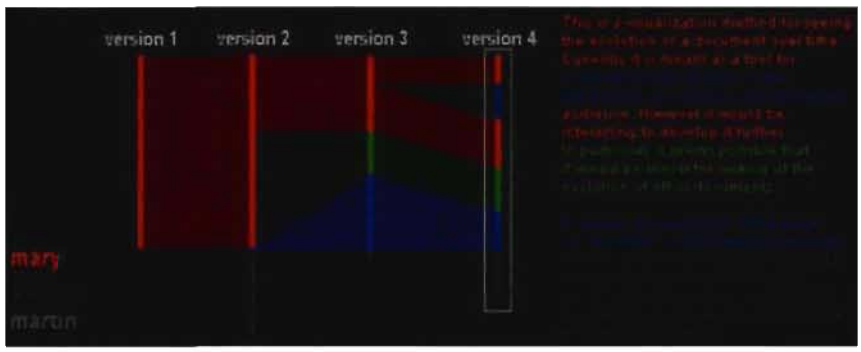

Figure 2.3 – Fonctionnement de history flow

Dans *history flow*, chaque version d'un article (ou page) est présentée par une barre verticale dont la taille est relative au nombre de caractères du texte de cette version. Ensuite la persistance d'un texte d'une version à la suivante est représentée par un quadrilatère (polygone à 4 côtés) dont deux côtes parallèles font partie des deux versions successives. De plus, chaque texte nouveau dans une version est associé à la couleur attribué à son auteur. Ainsi les couleurs de la barre verticale correspondent aux couleurs attribuées aux auteurs qui ont rajouté le texte pour la première fois. L'image ci-dessus illustre le fonctionnement de ce logiciel. Nous pouvons ainsi constater que :

- la première version a été écrite par *Mary* à qui le logiciel a associé la couleur rouge ;

- la première version a été modifiée par *Suzann* qui a rajouté un texte à la fin de l'article. Le nouveau texte ajouté est en bleu, la couleur associé à *Suzann*. On peut également constater que le texte rajouté par *Mary*, dans la première version est restée intacte dans la nouvelle version et ceci est marqué par le remplissage du quadrilatère correspondant avec la couleur rouge ;
- la deuxième version a été modifiée par *Martin* pour donner une troisième version de l'article. Dans cette dernière, nous pouvons constater que seulement une partie du texte introduit par *Mary* lors de la première version, est restée intacte. À la place de l'autre partie du texte, *Martin* a rajouté un autre texte dont la couleur est verte. Le texte rajouté par *Suzann* dans la deuxième version a persisté (couleur bleu du quadrilatère correspondant), ainsi de suite.

Les expérimentations de cet outil sur 70 articles de l'encyclopédie Wikipedia ont permit de dégager quatre principaux *scénarios collaboratifs* à savoir :

- *Vandalisme - Restauration.* Les chercheurs ont constaté que le vandalisme est exprimé de différentes façons : suppression totale du contenu d'un article, insertion de mots vulgaires et insultes, insertion d'un faux contenu sans relation avec le thème, insertion de fausses redirections ou encore insertion d'un contenu partial contrevenant à la $NPOV$[3]. La recherche a permis également de constater qu'en général les articles qui subissent ce genre d'abus, sont restaurés en un temps inférieur à trois minutes.
- *Négociation* ou ce qu'on appelle la *guerre d'édition* dans le jargon wikiste [24]. Il s'agit d'un phénomène durant lequel deux ou plusieurs éditeurs expriment un profond désaccord sur un point particulier (le contenu ou le titre d'un article, sa subdivision en plusieurs articles de petite taille, un paragraphe non neutre, etc.). La négociation se manifeste, entre autres par un cycle de révocation actif sur l'article lui-même (alternance d'ajout et suppression du même texte). C'est un scénario qui apparaît le plus souvent dans les articles controversés comme le Christianisme, l'Islam, etc.

[3]Neutralité de point de vue

- *Anonymat - Signature.* L'anonymat est la modification d'un article par des utilisateurs non-identifiés, seul leur adresse IP est connu. Tandis que la signature est la modification par un utilisateur enregistré. Les chercheurs ont notamment constaté qu'il n'y avait pas de lien entre la signature et le volume des contributions, ni de lien entre l'anonymat et le vandalisme.
- *Stabilité* ou la persistance du contenu. Certaines parties des articles ne persistent pas mais les expériences montrent que ceci est aléatoire dans le temps.

Durant notre manipulation du logiciel *history flow*, nous avons remarqué deux limitations :

1. Tout d'abord, le logiciel ne permet pas de visualiser de façon systématique, l'enchaînement des interventions à diverses niveaux de granularité. Or, il serait intéressant de pouvoir se focaliser sur une seule partie de l'article comme un paragraphe (qui en linguistique est la partie du texte contenant une idée et son développement), de pouvoir visualiser son évolution et détecter des scénarios collaboratifs plus spécifiques.
2. L'aspect sémantique des interventions n'a pas été considéré. En effet *history flow* met en évidence uniquement l'aspect syntaxique des interventions autrement-dit le type d'opération de modification ainsi que la taille du texte ajouté ou supprimé. Néanmoins, pour bien analyser la collaboration des utilisateurs sur les Wikis, il serait intéressant de voir le sens ou l'aspect sémantique des interventions (complétion, opposition, etc.).

2.2.1.2 Rigueur et diversité comme métriques d'évaluation

A. Lih [11] s'est intéressé à la corrélation entre la citation des articles de Wikipedia dans la presse et leur qualité. En se basant sur deux métriques pour évaluer la qualité d'un article : le nombre de révisions (ou versions) d'un article - qu'il appelle *rigueur* - et le nombre de contributeurs uniques - qu'il appelle *diversité* -, il a remarqué qu'un article cité dans la presse bénéficie automatiquement de cette

citation et gagne en qualité (plus d'interventions faites par différents utilisateurs). En effet, il a considéré les valeurs (rigueur et diversité) d'un certain nombre d'articles de Wikipedia avant et après leur citation dans la presse et il les a comparées avec celles d'un ensemble d'articles de référence (articles non cités dans la presse et susceptibles d'être complets et de bonne qualité : articles sur des sujets d'actualité et non controversés). Il a remarqué que les valeurs de certains des articles cités ont dépassé la valeur moyenne des articles de référence ce qu'il considère comme un gain de qualité. La conclusion de cette étude est néanmoins discutable. En effet, l'auteur a omis la justification de certains points de la méthodologie comme le fait qu'un grand nombre de révisions et un grand nombre de contributeurs uniques implique une meilleure qualité d'un article.

2.2.1.3 Évaluation de la qualité d'information par les pairs

Les deux scandales qui ont frappé l'encyclopédie collaborative vers la fin de l'année 2005 - la dénonciation d'un article de Wikipedia qui suggérait faussement que John Seigenthaler, un ancien assistant administratif de Robert Kennedy, aurait pu être impliqué dans les assassinats de Robert Kennedy et John F. Kennedy et, quelques jours plus tard, l'accusation contre Adam Curry, un pionnier états-uniens du podcasting, d'avoir édité un article sur ce nouveau média en enlevant les références sur les travaux des autres compétiteurs - ont remis en question la qualité des informations sur cette encyclopédie participative gratuite et ont, entre autres déclenché une étude [14, 15] *peer review* comparative de Wikipedia et de la vénérable encyclopédie Britanica.

Mené par la revue *Nature*, cette étude a révélé que les articles de Wikipedia ne sont guère moins fiables que ceux produits par Britanica. En effet, sur les 42 articles scientifiques analysés par des experts, 4 erreurs graves ont été identifiées respectivement dans les deux encyclopédies ainsi qu'un certain nombre d'inexactitudes (123 dans Britanica contre 162 dans Wikipedia). Le résultat final d'une moyenne de 2.92 erreurs par article de Britanica contre 3.86 pour Wikipedia est cependant

controversé : alors que Wikipedia se réjouit de cette étude qui montre finalement que les deux scandales cités plus-haut n'étaient qu'une exception et pas une règle, Britanica remet en question le processus utilisé par la revue Nature.

2.2.1.4 Modèle d'évaluation automatique de la qualité d'information

À ces études qui s'attachent à l'évaluation de la qualité des contenus des sites Wikis, il faut rajouter le travail de *Stvilia et al.* [20] dans lequel les auteurs ont proposé un modèle d'évaluation automatique de la qualité de l'information de Wikipedia. Ils ont considéré un certain nombre de caractéristiques statistiques des articles, entre autres le nombre de révisions, le nombre de liens, le nombre d'images, l'âge et la taille, qu'ils ont regroupés avec une certaine pondération pour chaque caractéristique pour former un ensemble de 7 métriques permettant de mesurer la qualité de l'information : *Authority/Reputation*, *Completeness*, *Complexity*, *Informativeness*, *Consistency*, *Currency* et *Volatility*.

Ces métriques ont été testées sur un corpus d'articles de la version anglaise de Wikipedia et la classification automatiquement de ces articles dans deux catégories : *Articles de Qualité*[4] ou *Featured Articles* et *non-AQ* donne de bons résultats. Bien que cette méthodologie est d'une grande utilité et efficacité (automatique) pour la communauté qui pourrait l'utiliser comme une classification préliminaire des articles, elle présente, comme l'ont souligné les auteurs, une lacune quant à l'évaluation de l'aspect sémantique des contenus.

Un élément critique qui n'a pas été considéré dans les études précédentes est la *cohérence textuelle* qui pourtant en linguistique est d'une grande importance lorsqu'on parle de la qualité d'un texte. Or, en observant les interventions des utilisateurs sur les sites Wikis, nous avons constaté que les auteurs cherchent à produire un

[4]Les AQ de Wikipedia sont une classe spécifique d'articles qui ont suivi un processus de reconnaissance de leur qualité par un vote de la communauté wikipédienne. Ils sont marqués par une étoile. http ://en.wikipedia.org/wiki/Wikipedia :Featured_articles

texte cohérent et par là, mettent explicitement les liens existant entre le contenu de leur intervention et le texte entourant. Ceci reflète, comme nous le verrons dans la suite de cette section, le principe linguistique de cohérence textuelle. Ce mémoire est donc un complément des travaux antérieurs par l'introduction des aspects sémantiques ou rôles des interventions comme la restriction, l'opposition, etc. Ce nouvel élément est par la suite utilisé pour identifier de nouveaux mécanismes de collaboration comme la jonction, un des procédés de la cohérence textuelle qui se manifeste entre autres, par l'utilisation des marqueurs.

2.2.2 Cohérence textuelle

Dans cette section, nous allons parler de la notion de cohérence textuelle. Après une introduction qui comprend la définition et les trois conditions de la cohérence textuelle, nous verrons en détail les différents procédés qui l'assurent.

2.2.2.1 Introduction

"Un texte cohérent est un texte dont on dit qu'il se tient, qu'il coule, dont on perçoit facilement l'unité malgré qu'il soit composé de plusieurs phrases" [16]. En revanche, un texte qui présente des coupures momentanées de l'unité, exigeant par ce fait que son lecteur revoie certains extraits ou reconstruise mentalement la chaîne de ses différentes phrases, sera considéré moins cohérent. Afin d'éviter que le lecteur se retrouve dans une telle situation, un texte doit respecter trois conditions suivantes :

- **la cohésion** : il s'agit de la *"qualité d'un ensemble dont les éléments paraissent reliés entre eux"*. Dans un texte, la cohésion est réalisée par des rappels, d'un énoncé au suivant, de ce dont on vient de parler autrement-dit la *cohésion thématique* et de la signification de ce que l'on vient d'en dire ce qui correspond à la *cohésion sémantique.*

  "`Jeanne et Alex sont partis pour une randonnée. Cependant ils ont oublié les chaussures de marche.`"

  Dans cet exemple, "ils" est un rappel thématique de "Jeanne et Alex" ; "chaus-

sures de marche" est un rappel à la fois thématique et sémantique de "randonnée" et "Cependant" indique le sens de la relation entre les deux phrases et contribue par ce fait à la continuité (ou unité) sémantique.

Plusieurs procédés comme la récurrence, la jonction, etc. permettent d'assurer la cohésion dans un texte, nous verrons plus en détails ces procédés dans la 2e partie de cette section.

- **La hiérarchisation** (des énoncés) : consiste en *"une indication de l'importance relative des énoncés ainsi qu'une inscription par là du point de vue privilégié d'où on se place pour développer une idée ou décrire un événement"*. Elle s'exprime notamment par l'ordre de présentation des énoncés : placement de l'énoncé principal en premier suivi des autres énoncés pour le développer (l'élaborer, l'expliquer ou l'illustrer). Cependant pour éviter qu'un texte soit vu comme une suite sans fin d'élaboration de l'énoncé principal, la *jonction* (ou la liaison par connecteurs) est utilisée pour ponctuer le texte et relancer un nouveau cycle "énoncé dominant (principal) - énoncé subordonnés". Dans l'exemple précédent, le connecteur "Cependant" joue ce rôle de ponctuation de texte tout en indiquant le sens de la relation entre les deux phrases. D'autres procédés de la hiérarchisation comme le regroupement, la coordination, etc. seront vus dans la partie consacrée aux procédés pour assurer la cohérence textuelle.
- **L'intégration** : *"exige que tout énoncé puisse être reconnu comme faisant partie du texte antérieur"*. Selon Pepin [16], le principe d'intégration est le plus important pour la cohérence. Cependant, contrairement aux deux autres conditions de la cohérence vues plus haut, celle-ci est caractérisée par un manque de procédés qui peuvent être appliqués pour l'assurer. Dans la partie qui suit, nous allons donc voir plus en détails les différents procédés pour assurer la cohésion et la hiérarchisation.

2.2.2.2 Procédés pour assurer la cohérence textuelle

La cohérence textuelle est un problème qui a intéressé un grand nombre de chercheurs[5] dans le domaine de la linguistique textuelle. Durant nos recherches, nous avons constaté que les procédés pour assurer la cohérence textuelle variaient suivant les auteurs. Dans ce mémoire, nous nous sommes limités aux procédés les plus cités et donc communs à plusieurs travaux [3, 4, 7, 12, 16].

2.2.2.2.1 La récurrence

La récurrence est une contrainte qui a été introduite pour la première fois par Bellert [3]. Elle consiste en une *"reprise d'éléments du texte dans son développement linéaire"*. Ce principe est aussi connu dans la littérature sous le nom de *règle de répétition* [4]. Les marques de ce procédé dans un texte sont de deux sortes :

- Les marques de reprise directe : elles indiquent l'*identité référentielle* entre les deux termes de la récurrence. Elles sont au nombre de quatre :
  - La *répétition*. "`Jeanne et Alex... Jeanne...`";
  - La *pronominalisation*. "`Le bus... Il...`";
  - La *définitivisation* (par déterminants définis ou démonstratifs). "`Un chat était poursuivi par un chien. Le chien... Ce chien...`";
  - La *substitution lexicale* (avec définitivisation). "`Jeanne s'est acheté une maison. Cet achat...`".

    Les trois dernières catégories se trouvent généralement classées dans la catégorie de procédés de la *coréférence : "substitution d'un élément du discours par un autre qui le représente"* [7]. Nous pouvons constater que dans les quatre cas cités ci-dessus, le recouvrement d'identité est total. Cependant on peut retrouver des exceptions où le recouvrement serait partiel. "`Les animaux... Certains...`".
- Les marques de reprise indirecte : elles indiquent non pas l'identité référentielle mais plutôt la *parenté référentielle*, autrement-dit l'appartenance

[5] Une bibliographie exhaustive dans ce domaine peut être trouvée ici : http ://www.philhist.uni-augsburg.de/lehrstuehle/anglistik/sprachwissenschaft/bibliography/

à un même champs sémantique. Il n'y a qu'une seule marque de reprise directe et c'est la *contiguïté sémantique* (avec déterminants définis ou possessifs) [12] appelé également *cohésion lexicale* [7]. "`Chat... Les pattes... Les griffes...`" Ceci pourrait être une suite de phrases où l'on parle des pattes du chat et de ses griffes.

Le rôle des récurrences est de participer au maintien de la continuité (ou unité) thématique du texte. Mais comme nous l'avons noté dans l'introduction sur la cohérence, l'auteur d'un texte doit également se préoccuper de la continuité sémantique, celle-ci est assurée par la jonction.

2.2.2.2.2 La jonction

"La jonction (ou liaison par connecteurs) est l'utilisation de connecteurs (car, cependant, mais, etc.) pour indiquer explicitement des relations sémantiques (cause, condition, opposition, etc.) entre des propositions, des phrases ou même des paragraphes" [7, 12, 16]. Ce procédé est également connu sous le nom de *règle de relation* [4].

Les connecteurs, connus aussi sous le nom de *marqueurs de relation* ou *organisateurs textuels* sont donc des mots qui aident à établir des liens (ou relations) entre différents éléments d'une phrase ou entre les parties d'un texte. Ils permettent d'identifier plus clairement chacune des parties du texte.

Dans [2] l'auteur a proposé une méthodologie de développement d'un ensemble de relations pouvant exister entre les parties d'un texte. En se basant sur un corpus de 350 connecteurs utilisés pour exprimer ces relations, il a formulé une classification (ou taxonomie) composé de 10 relations (*Sequence, Cause, Result, Restatement, Temporal, Negative polarity, Additional information, Hypothetical, Similarity et Digression*). D'autres classifications de ces relations sémantiques ont été proposées dont une qu'on peut dire plus générale et qu'on retrouve dans le livre

de Fairclough [5]. Cette classification générale comprend six (6) relations sémantiques (*Causal, Conditional, Temporal, Additive, Elaborative et Contrastive*). Le tableau 2.1 à la page 34 montre une mise en parallèle de ces deux classifications.

Relations sémantiques selon Alistair [2]	**Relations sémantiques générales [5]**	**Déscription**
Sequence	Additive	**ajout** d'un nouvel élément ou coordination de deux ou plusieurs
Cause Result	Causal	**cause** ou explication de ce qui cause l'idée émise, conséquence ou but
Hypothetical	Conditional	exprimer une **condition**, une supposition ou une hypothèse
Negative polarity	Contrastive	exprimer une **opposition**, une **concession** ou une **restriction**
Restatement Additional Information Similarity Digression	Elaborative	expliquer, affirmer, illustrer ou résumer une idée
Temporal	Temporal	situer le **moment** ou la **durée** d'une idée

Tableau 2.1 – Relations sémantiques

La textualité (ou l'écriture d'un texte cohérent) compte deux grands principes : la continuité (thématique et sémantique) et la progression [12]. En effet, lors de l'écriture d'un texte, l'auteur maintient l'unité ou la continuité du texte en préservant des éléments stables ou récurrents et fait progresser le discours en ajoutant des éléments nouveaux. De plus, comme nous allons le voir dans la partie qui suit, l'information nouvelle ne peut être intégrée à l'ensemble (à l'unité) que si elle est portée par des éléments connus.

2.2.2.2.3 La progression

Dans une phrase l'élément connu est appelé *thème* et les éléments nouveaux, *rhèmes* [16]. De manière générale, le thème est positionné en tête de la phrase et représente ce dont on parle tandis que le rhème énonce ce que l'on en dit. Le jeu d'articulation mis en oeuvre par les thèmes et les rhèmes, entre deux phrases, doit former un ensemble cohérent et est connu sous le nom de ***progressions thématiques***. D'après la typologie de Danes (1974 : voir [16]), il existe trois schémas principaux de progressions :

- les *progressions à thème constant.* Dans "`Jeanne s'est acheté un ordinateur. Elle l'a payé 800$`", les deux phrases ont le même thème ("Elle" est une substitution de "Jeanne") ;
- les *progressions linéaires.* Dans "`Jeanne s'est acheté un ordinateur. Il lui a coûté 800$`", le thème de la deuxième phrase (Il) est une substitution du rhème de la première phrase (ordinateur) ;
- les *progressions dérivées d'un hyper-thème*[6] *ou d'un hyper-rhème.* Dans "`Jeanne s'est acheté un ordinateur. le disque dur est de 80GB. l'écran est un 17 pouces.`", les deux derniers thèmes (Le disque dur) et (L'écran) sont dérivés de l'hyper-rhème de la première phrase.

2.2.2.2.4 La non-contradiction

Selon M. *Charolles* [4] *"le développement d'un texte cohérent ne doit introduire aucun élément sémantique contredisant un contenu posé ou présupposé par une occurrence antérieure ou déductible de celle-ci par inférence".* C'est une notion qui date de l'époque d'Aristote avec ce principe qu'*on ne peut en même temps affirmer une chose et son contraire.*

Comme exemple de ce procédé, on retrouve souvent cet anecdote d'un apprenti-logicien auquel son voisin avait prêté un chaudron et qui le lui avait rendu percé. À ces plaintes, l'apprenti-logicien répondit : "Premièrement, j'ai **rendu le chaudron**

[6] L'hyper-thème (hyper-rhème) est le thème (rhème) général, lui-même divisé en plusieurs thèmes qui, chacun d'eux, peut comprendre un ou plusieurs rhèmes

intact ; deuxièmement, il était déjà **percé au moment où je l'ai emprunté** ; troisièmement, je n'ai **jamais emprunté de chaudron**."

2.2.2.2.5 L'intertextualité

Le terme intertextualité est attribué à Julia Kristeva et réfère au caractère et à l'étude de l'intertexte (mise en relation de plusieurs textes par le biais notamment de la citation, la référence, etc.). H. Plett (1991)[7] distingue deux types d'intertexualité selon la façon d'assurer la cohérence textuelle : *l'intertextualité intratextuelle* qui garantit l'intégrité du texte et *l'intertextualité intertextuelle* qui lie structurellement le texte (intertexte) à d'autres textes. Les marques de ce procédé dans un texte sont : la citation, l'allusion, la réécriture ou le remaniement.

Pour qu'un texte soit cohérent, les procédés pour assurer la cohérence textuelle que nous venons d'énumérer doivent être **régulièrement** et **suffisamment** présentés dans ce texte [16]. Nous soutenons que lors de la modification des contenus, les wikistes n'échappent pas à cette règle, particulièrement lorsqu'ils veulent marquer les relations sémantiques entre leurs idées et celles émises auparavant (le procédé de jonction).

Afin d'identifier les mécanismes de collaboration comme le maintient de la cohérence sur un site Wiki, tout en tenant compte de cet aspect **fréquentiel minimal** des procédés pour assurer la cohérence ainsi que du caractère **temporel** des interventions sur les pages d'un Wiki (différentes versions), nous avons eu recours à une technique de fouille de données : les motifs séquentiels, pour dégager des séquences d'interventions fréquentes (relativement à un seuil fixé *à priori*), manifestations de ces mécanismes de collaboration, dans un corpus d'historiques des modifications d'un site Wiki. La suite de cette section présente les notions de base de fouille de donnée et d'extraction de motifs fréquents.

[7] http ://public.enst-bretagne.fr/~thliviti/th/node14.html

2.2.3 Fouille de données et extraction de motifs fréquents

Face à l'immense quantité de données qu'elles collectent et stockent tous les jours, il s'est avéré vital pour les entreprises, de disposer de techniques pour analyser ces données afin d'y extraire des informations ou connaissances de façon automatique. Ce processus est connu sous le nom de *découverte de connaissances dans les bases de données* (Knowledge discovery in Databases ou KDD) et s'effectue en 5 étapes. Tout d'abord les données sont **sélectionnées**, le cas échéant **pré-traitées**, et **transformées** pour faire l'entrée du processus de **fouille de données**. Le résultat de cette dernière étape constitue les nouvelles connaissances recherchées. Elles sont enfin visualisées afin d'être **interprétées** et **évaluées**.

Cette section est divisée en trois parties : la première partie donne quelques notions de base de la fouille de données notamment sa définition, quelques-unes de ses applications et une exposition des différentes tâches de ce processus. Dans la deuxième partie, nous nous focaliserons sur l'extraction de motifs fréquents, une étape commune à certaines tâches de la fouille de données (notamment les règles d'association et sa dérivée les motifs séquentiels) et nous verrons un peu plus en détail l'approche *Apriori* pour l'extraction de motifs fréquents. Nous finirons cette section par notre adaptation du processus de KDD à notre approche.

2.2.3.1 Fouille de données

"La fouille de données est le processus non trivial d'extraction de connaissances implicites, précédemment inconnues et potentiellement utiles à partir de données." [9]. Elle permet d'extraire plusieurs formes de connaissances à partir de données : soit sous forme de règles, de modèles, de régularités, de concepts, etc. Ceci rend le domaine de fouille de données très actif et lui procure un large champ d'applications à divers domaines tels que :

- le marketing et la vente : les transactions commerciales des clients peuvent être analysées afin d'optimiser des réapprovisionnements. Cette analyse est

également connu dans la communauté sous le terme d'*analyse du panier de la ménagère.*

- la finance, l'assurance : l'analyse des données personnelles (sexe, age, profession, etc.) ou celle des données sur les éléments à assurer (type de voiture, taille de maison, etc.) permettent aux sociétés d'éliminer des "mauvais" clients, d'évaluer des risques, d'autoriser de crédits aux "bons" clients, de détecter des fraudes, de proposer des services spécifiques aux clients, etc.
- l'informatique : les techniques de la fouille de données permettent notamment d'analyser automatiquement du courrier électronique et de rejeter des messages susceptibles de contenir des virus, elles sont également utilisées dans l'analyse automatique de sites Web par des certains moteurs de recherche, etc.

2.2.3.1.1 Approches principales de la fouille de données

L'abondance des applications de la fouille de données sous-entend une multitude de données à analyser : base de données relationnelles, base de données transactionnelles, base de données orientées objets, base de données relationnelles objets, base de données temporelles (exemple : la bourse), base de données spatiales (exemples : Images provenant de satellites, cartes géographiques), base de données textuelles (exemples : le Web, le courrier électronique, les pages html/xml), etc.

Cette variété de données explique d'une part l'existence de la deuxième et troisième étape du processus de KDD et d'autre part le besoin de techniques spécifiques de fouille de données. Durant nos recherches, nous avons constaté que le nombre de ces techniques varie selon les auteurs, dans la liste ci-dessous nous nous sommes focalisés à celles qui reviennent le plus souvent dans la littérature.

1. ***Classification et prédiction*** : permet la division ou groupement d'instances dans des classes spécifiques suivant un ensemble de prédicats les caractérisant. Ces derniers peuvent être par après appliqués à des objets inconnus afin de prévoir leur classe d'appartenance.
   Exemple : classement des clients par une banque en deux classes : les clients loyaux / les clients non-loyaux, afin d'accorder un crédit.
   Plusieurs algorithmes permettent de mettre en oeuvre cette technique : arbres de décision, règles de classification, classification Bayésienne, algorithmes génétiques, algorithme des k plus proches voisins, l'approche Rough Sets, régression linéaire et non linéaire.
2. ***Regroupement (Clustering)*** : permet le regroupement d'éléments de proche en proche fondé sur leur ressemblance. Contrairement à la précédente approche, les classes sont inconnues et sont donc créées.
   Parmi les algorithmes utilisés pour cette approche citons : K-moyennes et réseaux neuronaux.
3. ***Règles d'association*** : permet la découverte de règles intelligibles et exploitables dans un ensemble de données volumineux, règles exprimant des associations entre *items* ou *attributs* dans une base de données [17]. Une règle d'association est une implication de la forme : $X \Rightarrow Y$, où X et Y sont deux ensembles d'items (ou *itemsets*) qui n'ont aucun item en commun. Elle est associée à deux notions importantes, à savoir, le ***support*** et la ***confiance***, qui permettent de l'évaluer.

   Le support d'une règle d'association $X \Rightarrow Y$ est noté $Supp(X \Rightarrow Y)$, c'est le nombre d'itemsets contenant à la fois les items de X et ceux de Y.
   La confiance d'une règle d'association $X \Rightarrow Y$ est noté $Conf(X \Rightarrow Y)$, c'est le nombre d'itemsets contenant les items de Y parmi les itemsets contenant les items de X.

Tableau 2.2 – Base de données contenant des transactions commerciales

TID	Client	Date	Transaction
1	cl1	2007-05-05	Télévision, lecteur DVD
2	cl2	2007-06-15	Télévision, home-cinéma, DVDs
3	cl3	2007-08-13	Télévision, home-cinéma
4	cl3	2007-08-27	DVDs
5	cl4	2007-12-23	Télévision, lecteur DVD, DVDs

Supposons que le tableau 2.2 à la page 40 illustre une base de données dans laquelle ont été enregistrées des transactions commerciales. On peut constater qu'il y a 2 personnes (cl2 et cl3) qui ont acheté une télévision et un home-cinéma en même temps. La règle d'association "télévision $\Rightarrow$ home-cinéma" peut ainsi être généré avec comme support $2/5 = 40\%$ et parmi les 4 transactions (TID=1,2,3,5) où les clients ont acheté une télévision, seules dans la deuxième et troisième transaction, les clients ont aussi achété un home-cinéma, la confiance de cette règle est donc de $2/4 = 50\%$.

Le processus d'extraction des règles d'association se fait généralement en deux étapes. Tout d'abord un ensemble d'itemsets fréquents ou motifs fréquents est extrait puis les règles d'association sont générées, dans un deuxième temps à partir de ces motifs.

Si l'on considère l'exemple ci-dessus, plusieurs itemsets peuvent être considérés : "télévision, lecteurs DVD", "télévision, home-cinéma, DVDs", etc. et leurs supports (ou les nombres de transactions dans la base de données contenant ces itemsets) peuvent être calculés afin de valider leur fréquence par rapport à un support de référence (appelé également support minimum). Ensuite les règles d'association relatives aux motifs peuvent être générées.

Dans ce mémoire, nous nous intéressons plus à la première étape, plus de détails sur son fonctionnement seront donnés dans la suite de ce mémoire. Les

algorithmes utilisés pour extraire des règles d'association diffèrent le plus souvent sur la façon dont la première étape est effectuée. Parmi ces algorithmes, nous pouvons citer : Apriori, Partition, D.I.C., SAMPLE, Closet, ect.

4. ***Motifs séquentiels*** : c'est une variation de la première étape des règles d'association qui prend en compte le temps entre les transactions des clients [1]. Toujours en considérant les transactions illustrées dans le tableau 2.2 à la page 40, on peut constater que le client cl3 a acheté une télévision et un ensemble home-cinéma en même temps, et qu'il est revenu deux semaines après pour acheter des DVD. Le motif séquentiel $\langle \{télévision, home\text{-}cinéma\} \{DVD\} \rangle$ peut donc être extrait avec un temps $t = 14 jours$ et son "support" – dans ce cas-ci le nombre de clients ayant manifesté ce même comportement – peut être calculé pour l'évaluer (dans ce cas-ci 1/4 ou 25%).
   Les algorithmes de cette approche sont généralement des variations de ceux utilisés dans le cas des associations : AprioriAll, GSP, PSP, etc.

2.2.3.2 L'extraction de motifs fréquents

L'extraction des itemsets ou motifs fréquents dans les bases de données est une phase primordiale dans le processus d'extraction des règles d'association, elle est commune à diverses autres tâches de fouille de données telles que l'extraction des corrélations, les motifs séquentiels, etc. C'est aussi une phase coûteuse en temps d'exécution car le nombre d'itemsets fréquents est exponentiel par rapport au nombre d'items dans la base de données. Plusieurs études dans le domaine se sont d'ailleurs focalisées sur l'amélioration de cette phase.

L'approche Apriori est la plus connue dans ce domaine car c'est la base de tous les autres algorithmes d'extraction de motifs fréquents. C'est cette approche que nous avons utilisée dans ce mémoire pour sa simplicité. La section suivante détaille son fonctionnement.

2.2.3.3 Principes de l'algorithme Apriori

Apriori est un algorithme itératif de recherche des itemsets (motifs) fréquents. Autrement-dit, durant la k-ème itération, un ensemble d'itemsets candidats C_k, avec k le nombre d'items dans l'itemset, est généré et un *scan* ou parcours de la base des transactions est réalisé afin de supprimer les candidats non fréquents. L'ensemble des k-itemsets fréquents ainsi généré est utilisé lors de l'itération $k+1$ suivante pour générer les candidats de taille $k+1$.
Ainsi, à chaque niveau ou itération, l'algorithme Apriori réduit l'espace de recherche en utilisant le principe suivant : *si un itemset de longueur k est non fréquent alors tous ses sur-ensembles (super-sets) le sont également.*

Plus particulièrement, l'extraction des motifs par l'approche Apriori consiste, pour chaque itération k, en une répétition des deux tâches :

1. La génération de l'ensemble des motifs candidats C_k ou ***jointure***. En effet, la génération des itemsets candidats de taille $k+1$ consiste en une jointure des itemsets fréquents de taille k. Les itemsets de taille 1 est un cas particulier et consiste tout simplement en des singletons des différents items de la base des transactions. Cette étape doit également satisfaire la condition suivante : un item est pris en compte une seule fois dans un même itemset. De ce fait, un 2-itemset candidat "DVD, DVD", généré par jointure de deux 1-itemset "DVD" et qui indique que le client a acheté deux DVD n'est pas pris en compte.
2. Le parcours de la base des transactions et élimination des itemsets candidats non-fréquents. Ceci consiste en un calcul du support des motifs candidats générés (ou le nombre des transactions de la base qui contient les même items que l'itemset candidat) et comparaison de celui-ci avec le support minimum $supp_{min}$ spécifié par l'utilisateur.

Ces deux étapes sont répétées jusqu'à ce que l'ensemble des itemsets ou motifs candidats C_k soit vide.

De façon générale, la phase de jointure consiste à rajouter un à un, les différents items, à chaque itemsets fréquents de l'étape précédente afin de générer les itemsets candidats.

Cependant, dans le cas d'une extraction de motifs séquentiels, cette opération d'ajout d'item est effectuée de deux manières : soit en ajoutant chaque item dans un itemset existant (exemple : $\langle\{\textit{télévision}, \textbf{\textit{home-cinéma}}\}\rangle$), ce qui renvient à dire que les différents items ont été achetés en même temps, soit en ajoutant l'item dans un nouveau 1-itemset (exemple : $\langle\{\textit{télévision}\}\{\textbf{\textit{home-cinéma}}\}\rangle$), ce qui revient à dire que cet item a été acheté quelque temps après.

Dans [19] les auteurs ont également proposé une variation des motifs séquentiels qui considère que les items peuvent faire partie d'une taxonomie (hiérarchie *est-une-sorte-de*). Ceci permet d'extraire des motifs contenant des items à différents niveaux de la taxonomie par l'application d'une opération de substitution. Ainsi à partir du motif $\langle\{\textit{télévision}\}\rangle$, on peut générer un motif plus spécifique $\langle\{\textit{télévision HD}\}\rangle$, car une "télévision HD" est une sorte de "télévision".

En résumé, dans le cas d'une extraction de motifs séquentiels où on considère que les items font partie d'une taxonomie, la génération de motifs candidats se fait en appliquant trois opérations aux motifs fréquents de l'étape d'avant, à savoir :

- ajout d'un item dans un itemset ;
- ajout d'un item dans un nouveau 1-itemset ;
- substitution d'un item par ses fils immédiats dans la taxonomie.

Bien que ces opérations constituent un mécanisme efficace de génération des motifs candidats [13], leur utilisation nécessite une précaution afin d'éviter la génération des doublons. En effet, en appliquant naïvement la première ou la troisième opération, un même itemset candidat peut être généré plusieurs fois. Par exemple, le motif $\langle\{\textit{télévision HD}, \textit{home-cinéma}\}\rangle$ peut être généré à partir du mo-

tif $\langle\{télévision, home\text{-}cinéma\}\rangle$ par substitution ou à partir de $\langle\{télévision\ HD\}\rangle$ par ajour d'un item.

Ce type de redondance peut être évité en fixant la position où les opérations vont être appliquées. Ceci renvient à appliquer des ***opérations canoniques*** ou opérations impliquant juste les derniers ensembles. Ainsi pour la première opération, il faut ajouter l'item dans le dernier itemset et pour la troisième opération, il faut substituer le dernier élément du dernier itemset par ses fils immédiats dans la taxonomie.

Enfin, dans le but d'obtenir des motifs fréquents intéressants, les données de la base à analyser doivent être normalisées. Par exemple, dans l'analyse des ventes dans une épicerie, on va considérer le type de l'article acheté (pourcentage de clients qui ont acheté des bières) au lieu de sa marque (pourcentage de clients qui ont acheté des bières "Molson dry" ou ceux qui ont acheté des bières "Heinekein"). Ceci revient, en quelques sortes, à associer des étiquettes aux articles (ou les catégoriser) et à considérer ces étiquettes (ou catégories) dans l'analyse des motifs fréquents plutôt que de considérer les articles.

2.2.3.4 Adaptation du processus de KDD à notre problématique

Pour rechercher les motifs de collaboration fréquents, nous nous sommes inspirés du processus de KDD dont nous venons d'introduire et nous avons adopté les différentes étapes de ce processus afin d'analyser les historiques des modifications des pages d'un site Wiki.

En effet, comme nous le verrons d'une manière plus détaillée dans le troisième chapitre de ce mémoire, la méthodologie proposée s'effectue en cinq étapes faisant référence aux cinq étapes de KDD mentionnées plus haut. Le tableau 2.3 à la page 45 résume cette adaptation.

Tableau 2.3 – Adaptation des étapes de KDD dans l'approche proposée

KDD	**Notre approche**
Sélection	Bases de données textuelles : historique(s) des modifications d'une ou plusieurs pages d'un site Wiki au format XML (voir le chapitre suivant)
Pré-traitement	Calcul de différence entre chaque paire de versions successives de la page afin d'identifier les interventions des utilisateurs et organisation de ces dernières en séquences pour respecter l'aspect temporel des versions (d'une page).
Transformation	Catégorisation des interventions (dans les séquences) suivant leurs caractéristiques syntaxiques et sémantiques. Ceci correspond à la normalisation des items (considération du type au lieu de la marque des articles achetés).
Fouille de données	Extraction des motifs séquentiels - des séquences de catégories d'interventions répétitives - en utilisant l'algorithme Apriori (l'item dans ce cas est une catégorie d'interventions).
Interprétation et évaluation	Visualisation des motifs afin de les évaluer, les interpréter et formuler des mécanismes de collaboration propres aux sites Wiki (par les spécialistes en communication).

CHAPITRE 3

NOTRE APPROCHE

Dans ce chapitre nous allons parler de façon détaillée de la méthodologie d'extraction de motifs de collaboration fréquents (ou séquences de catégories d'interventions répétitives) que nous avons proposée afin d'aider les chercheurs en Communication de l'Université de Montréal à identifier les mécanismes de collaboration qui émergent sur des sites Wiki. Pour ce faire, nous présenterons en premier lieu l'architecture de cette méthodologie puis nous verrons plus en détail le fonctionnement des cinq étapes composant cette architecture.

3.1 Architecture

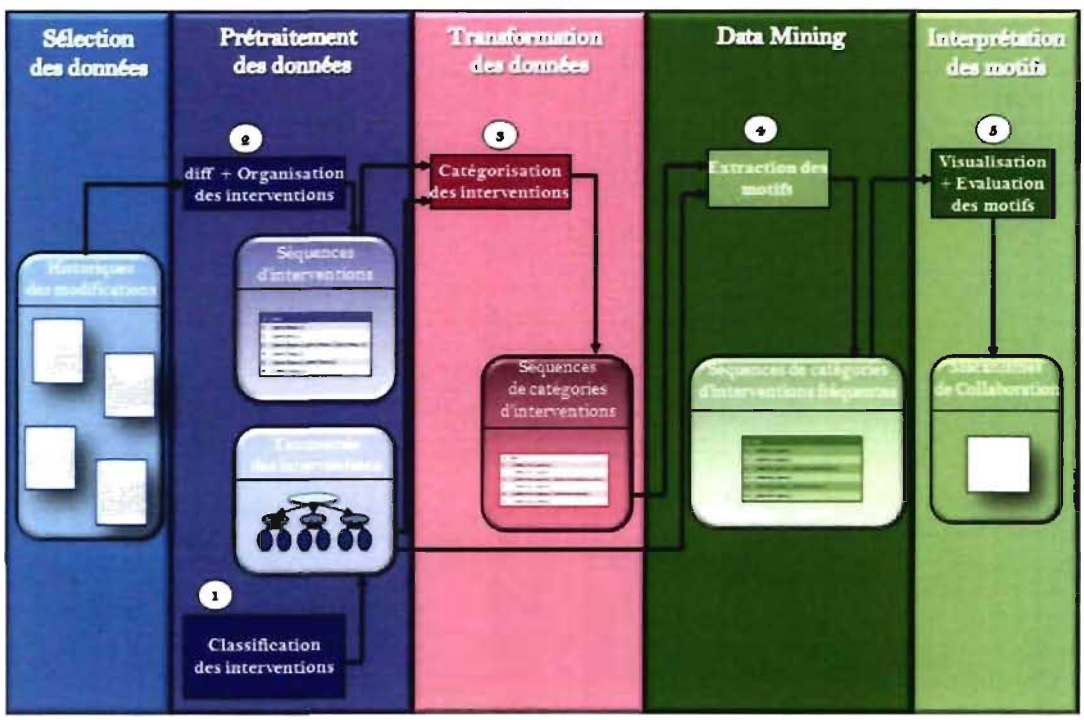

Figure 3.1 – Architecture de l'approche proposée

Comme nous l'avons mentionné, la problématique de recherche de motifs de collaboration fréquents est une situation de la fouille de données et, plus généralement, fait partie du processus KDD. Ceci est reflété sur l'architecture de notre méthodologie comme nous pouvons le constater sur la figure 3.1 à la page 46. En effet, nous avons proposé une approche en 5 étapes, une adaptation des 5 étapes du processus de KDD.

Dans cette architecture, chaque bloc (différente couleur) correspond à une étape de KDD. Tout d'abord la sélection des données. Dans notre cas, nous manipulons des historiques des modifications des pages d'un site Wiki (en l'occurrence les logs des articles de Wikipedia). La figure 3.2 à la page 47 donne un aperçu de l'historique des modifications de l'article "Université de Montréal" de Wikipedia.

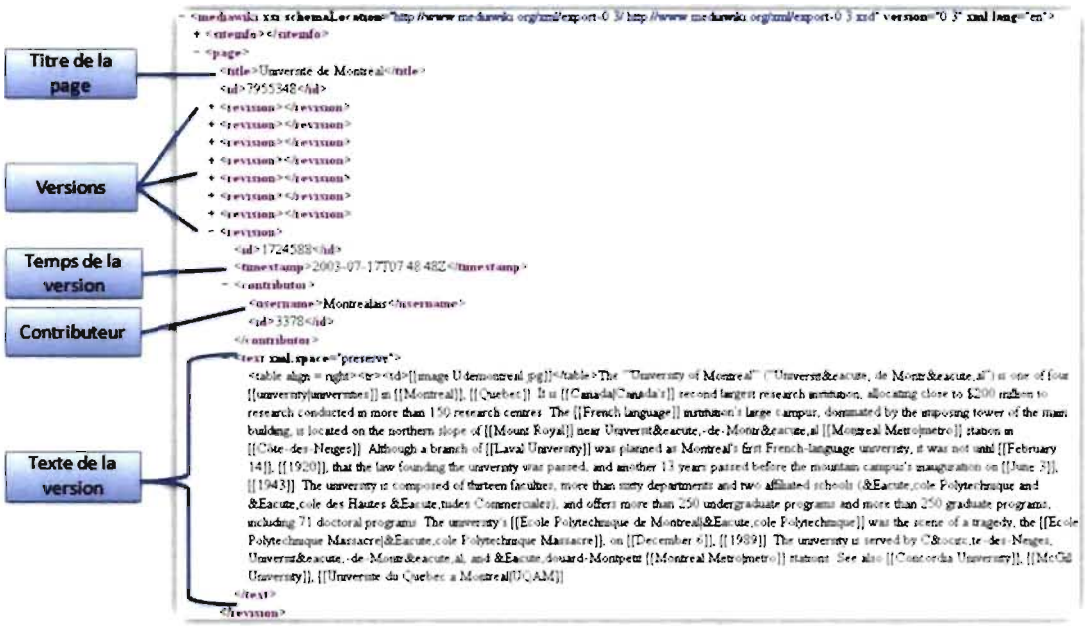

Figure 3.2 – Université de Montréal - Historique des modifications (Wikipedia)

Une fois les logs sélectionnés, la deuxième étape est consacrée à leur prétraitement. Cette étape est subdivisée, dans notre cas, en deux tâches. D'une part la classification des interventions en s'inspirant notamment des procédés pour assurer la

cohérence textuelle et d'autre part l'organisation des interventions venant d'un corpus de logs (d'articles de Wikipedia) dans des séquences.

Les séquences d'interventions obtenues durant la deuxième étape sont transformées en séquences de catégories d'interventions afin de permettre l'extraction de motifs fréquents intéressants. En effet les interventions contiennent un texte propre à une page particulière de Wikipedia ; afin d'extraire des motifs représentatifs des différentes pages, il faut considérer des données communes à plusieurs articles.
Pour cela, il faut **considérer les catégories d'interventions auxquelles appartiennent les interventions à la place des interventions elles-même.** La taxonomie des interventions constituée durant la deuxième phase est donc utilisée ici afin de catégoriser chaque intervention.

Les différentes séquences de catégories d'interventions constituées précédemment sont ensuite fouillées afin d'y extraire des séquences de catégories d'interventions fréquentes. La dernière phase de notre architecture consiste alors à présenter les motifs de collaboration fréquents à l'utilisateur (en l'occurrence les chercheurs en communication) afin qu'il puise interpréter ces derniers.

Dans la section qui suit nous allons détailler le fonctionnement des différentes tâches marquées de 1 à 5 sur l'architecture.

3.2 Fonctionnement

3.2.1 Classification des interventions

Comme nous l'avons indiqué dans la section consacrée aux définitions du chapitre 2, les interventions des utilisateurs lors de la modification des pages d'un site Wiki peuvent être classées suivant deux classifications (syntaxique et sémantique).

En effet, supposons que l'on ait identifié un grand nombre d'interventions ayant

consisté en l'ajout d'un texte dont le nombre de caractères dépasse un certain seuil et contenant au début, les mots-clés comme "Unfortunately, However, etc." (la figure 3.3 à la page 49 illustre ce type d'interventions).

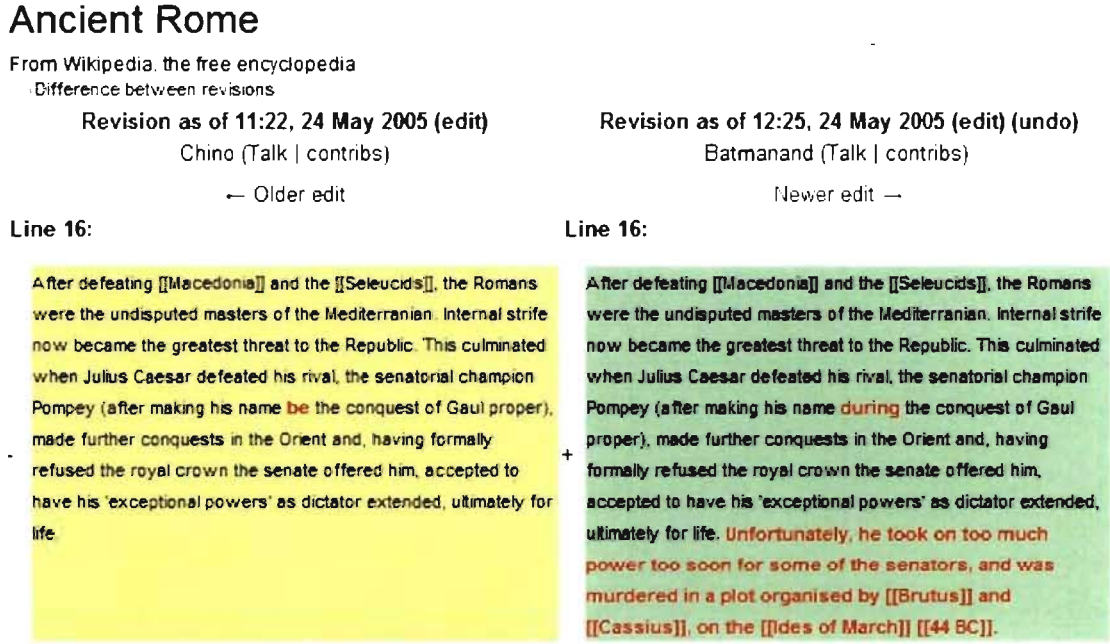

Figure 3.3 – Ajout d'un long texte pour exprimer une limite à une idée émise

Suivant une *classification syntaxique*, ces interventions peuvent être classées dans la classe "Ajout d'un long texte" et suivant une *classification sémantique*, elles peuvent être classées dans la classe "Restriction" car l'utilisation de ces mots-clés sous-entend en linguistique que l'on veut exprimer une limite à une idée ou énoncé émis. La description générale de ces interventions serait alors "Ajout d'un long texte pour exprimer une limite à l'idée émise" correspondant à une catégorie d'une taxonomie des interventions.

Sur la base de différents aspects des interventions observés avec le logiciel *historyDiff*, nous avons pu former une taxonomie décrivant les interventions sur un site Wiki. En effet, cet outil dont le principal objectif était de visualiser la différence

entre diverses versions d'une page sur un site Wiki (ou l'historique des modifications), nous a permis d'observer plusieurs interventions issues d'un corpus d'articles de Wikipedia et de dégager différents points communs à ces dernières. À partir de ces caractéristiques, nous avons pu constituer deux classifications des interventions et en déduire une taxonomie des interventions.

Dans la suite, nous allons nous intéresser aux différentes valeurs de ces deux classifications (ou taxonomies) syntaxique et sémantique. Par après, nous allons voir l'élaboration d'une taxonomie des interventions à partir de ces classifications.

3.2.1.1 La classification syntaxique

Pour constituer la taxonomie syntaxique représentant les interventions des utilisateurs sur un site Wiki, nous nous sommes intéressés à deux aspects dont le premier est le **type d'opération de modification** des interventions. En effet, nous avons constaté qu'une première classification des interventions pouvait se faire en considérant les deux types d'opérations de modification de base, à savoir, **l'insertion d'un nouveau texte** et **la suppression d'un texte**.

Bien que lors de la visualisation d'un corpus d'historiques des modifications avec le logiciel *historyDiff*, nous avons observé d'autres opérations comme le déplacement et le remplacement d'un texte, ces derniers éléments n'ont pas été considérés dans la taxonomie car ils peuvent être vus comme étant une combinaison des deux opérations de base considérées.

L'autre aspect qui nous a préoccupé est la **taille du texte** d'une intervention autrement-dit le nombre de caractères du texte ajouté ou supprimé. Cet élément nous a permis d'enrichir la classification syntaxe précédemment formée. Pour ça, nous avons scindé en deux classes plus spécifiques, chaque type d'opération considéré auparavant, suivant la taille du texte considéré : soit long (si le nombre de caractères est supérieur à x caractères. x étant un paramètre) ou court (si le nombre

de caractères est inférieur ou égal à x caractères).

La figure 3.4 à la page 51 illustre la taxonomie syntaxique des interventions résultante. Il s'agit d'une classification à **deux niveaux** et dont les éléments sont reliés par la relation "is-a" ou "est une sorte de". Le tableau 3.1 à la page 51 qui suit cette taxonomie donne les détails sur chaque élément composant la taxonomie.

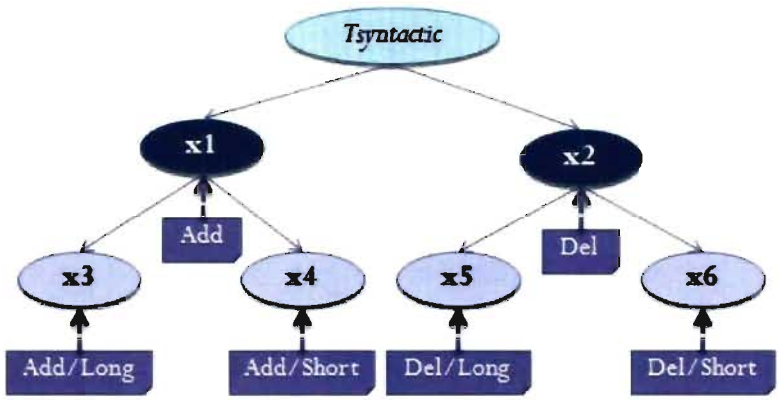

Figure 3.4 – Classification syntaxique des interventions

xi	Valeur/description
x1	**Add** : classe qui regroupe de façon générale toutes les interventions ayant consisté en l'ajout d'un texte (opération de modification est $insert(t_x)$).
x2	**Del** : classe qui regroupe de façon générale toutes les interventions ayant consisté en la suppression d'un texte (opération de modification est $delete(t_x)$).
x3	**Add/Long** : classe qui regroupe toutes les interventions dont l'opération de modification est $insert(t_x)$ et dont la taille du texte est $>$ à x caractères.
x4	**Add/Short** : classe qui regroupe toutes les interventions dont l'opération de modification est $insert(t_x)$ et dont la taille du texte est $\leq$ à x caractères.
x5	**Del/Long** : classe qui regroupe toutes les interventions dont l'opération de modification est $delete(t_x)$ et dont la taille du texte est $>$ à x caractères.
x6	**Del/Short** : classe qui regroupe toutes les interventions dont l'opération de modification est $delete(t_x)$ et dont la taille du texte est $\leq$ à x caractères.

Tableau 3.1 – Les valeurs de la taxonomie syntaxique

3.2.1.2 La classification sémantique

Le type d'opération de modification ainsi que la taille du texte (ajouté ou supprimé) dans une intervention apportent peu quant au sens de cette intervention. Afin d'enrichir cet aspect, nous nous sommes intéressés aux relations sémantiques existant entre les idées d'une intervention et celles du texte l'entourant.

Pour ce fait, nous nous sommes inspirés du procédé de jonction pour assurer la cohérence textuelle vu plus haut et nous avons constitué une classification sémantique des interventions basée sur les différentes relations sémantiques (marqué par différents marqueurs de sens ou connecteurs) exprimées par les wikistes lors d'une intervention sur une page.

Dans un premier temps, nous avons considéré la liste de connecteurs proposés par Alistair [2] et qui, selon lui, représentent de **façon exclusive** chacune des dix relations sémantiques qu'il a également proposées. Nous avons ensuite classé ces connecteurs dans les six relations sémantiques générales suivant la correspondance entre les deux corpus de relations sémantiques proposée dans le chapitre précédent (voir le tableau 2.1 à la page 34).

Le choix du corpus de connecteurs proposés par Alistair vient du fait que nous comptions évaluer notre méthodologie sur des articles venant de la version anglaise de Wikipedia mais surtout parce que c'est de loin le plus complet corpus de connecteurs en anglais que nous avons trouvé durant nos recherches. Le tableau 3.2 à la page 53 présente les six relations sémantiques et les connecteurs exclusifs correspondants.

Relation	Connecteurs
1. Additive	besides, first, first of all, firstly, for a start, for another thing, for one thing, furthermore, in addition, lastly, likewise, moreover, next, on top of this, secondly, thirdly, to begin with, to start with, what is more
2. Causal	accordingly, as a consequence, as a result, at once, at that, because, clearly, consequently, considering that, for, given that, hence, immediatly, in case, in doing this, in order that, in so doing, in that, insofar as, instantly, it follows that, now, now that, obviously, on the grounds that, plainly, seeing as, so, so that, thereby, therefore, this implies that, this way, thus, to the extent that, to this end
3. Conditional	as long as, assuming that, if, if ever, if only, if so, in that case, on condition that, on the assumption that, suppose that, supposing that
4. Contrastive	all the same, although, but, despite this, even so, even though, having said that, however, if not, in spite of this, instead, nevertheless, nonetheless, on one hand, on the one hand, otherwise, rather, still, then again, though, unless, whereas, yet
5. Elaborative	actually, as a matter of fact, at any rate, at least, by the way, e.g., even, for example, for instance, incidentally, in actual fact, in fact, in point of fact, in truth, indeed, just as, on the contrary, or rather, summing up, the way, to recap, to sum up, to summurise
6. Temporal	beforehand, ever since, meanwhile, previously

Tableau 3.2 – Connecteurs exclusifs et relations sémantiques exprimées

Une première classification sémantique des interventions basée sur ces six relations a donc pu être formée. En effet, nous supposons que les wikistes utilisent les différents marqueurs lors de la modification du texte afin d'exprimer de façon explicite le sens de leurs interventions et que ces dernières peuvent donc être classées suivant cet aspect sémantique. Néanmoins, il arrive parfois que le sens de l'intervention soit implicite (pas de marqueur) ou difficile à détecter de façon automatique (connecteurs exprimant plusieurs relations) et il fallait avoir une catégorie regroupant toutes les interventions de ce genre. La classification sémantique finale ainsi formée est illustrée sur la figure 3.5 à la page 54.

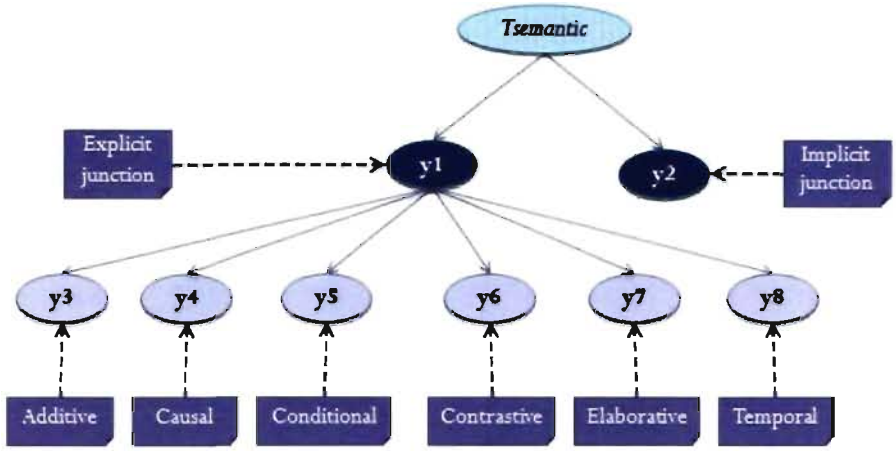

Figure 3.5 – Classification sémantique des interventions

3.2.1.3 Taxonomie des interventions

Après la constitution des deux classifications (syntaxique et sémantique), nous nous sommes intéressés à l'élaboration d'une taxonomie rassemblant les deux aspects.

Pour ce faire, nous avons accouplé les différentes valeurs des deux classifications formées précédemment. Les éléments de cette taxonomie sont appelés, comme nous l'avons défini plus haut, des catégories d'interventions.

Tout d'abord, une première catégorie d'interventions T correspondant à la racine de la taxonomie est générée. Ensuite, le premier niveau de la taxonomie est généré et comprend tous les couples (a, b) dont a est un fils immédiat de *tsyntactic* dans la classification syntaxique (autrement-dit $x1$ ou $x2$) et b est un fils immédiat de *tsemantic* dans la classification sémantique (autrement-dit $y1$ ou $y2$). Par la suite, la génération d'une *catégorie fils* (pour les niveaux 2 et 3) se fait en appliquant deux opérations à la *catégorie père*. D'une part, la spécialisation de sa valeur syntaxique (ou remplacement par les successeurs immédiats dans la classification syntaxique) et d'autre part, la spécialisation de sa valeur sémantique (ou remplacement de sa

valeur sémantique par les successeurs immédiats dans la classification sémantique).

La figure 3.6 à la page 55 illustre la construction de cette taxonomie. Nous pouvons ainsi constater que la catégorie $c5$ est obtenu en spécialisant $x1$ de $c1$ autrement-dit en le remplaçant par $x3$.

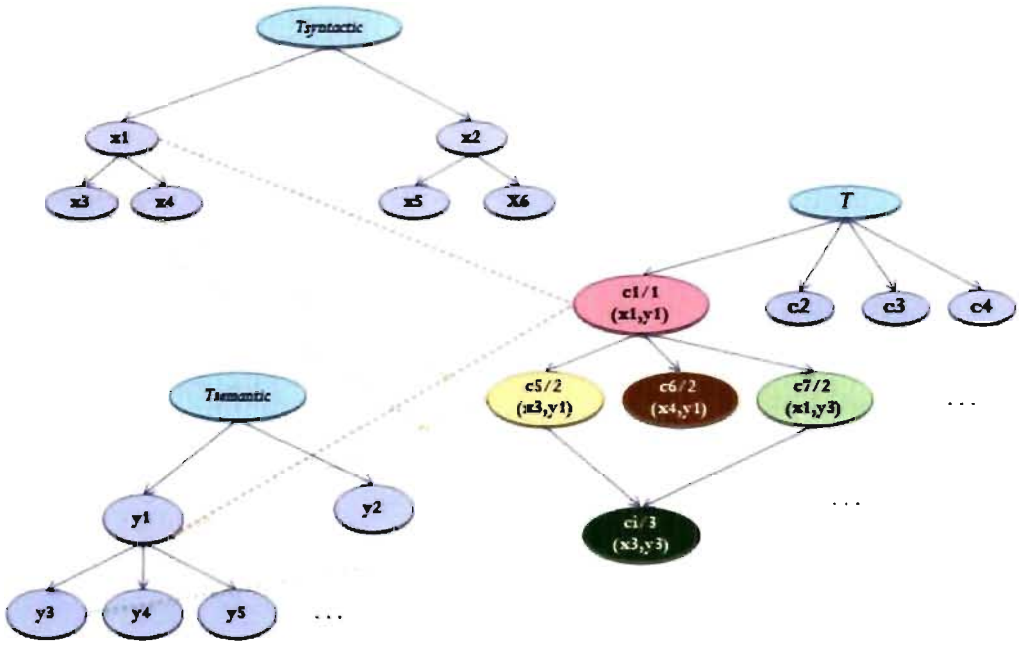

Figure 3.6 – Construction de la taxonomie des interventions

La taxonomie ainsi formée contient 48 catégories d'interventions de trois types, suivant leur positionnement :

- la *racine* ;
- les *catégories intermédiaires* : les catégories possédant des fils (pas la racine) ;
- les *catégories feuilles* : les catégories ne possédant pas de fils.

Le tableau 3.3 à la page 56 contient les 28 catégories feuilles ainsi que leurs interprétations.

Catégorie texte court	Catégorie texte long	Description
C13	C14	Ajout d'un texte sans exprimer la relation sémantique
C23	C24	Suppression d'un texte sans marqueur pour exprimer la relation sémantique
C25	C31	Ajout d'un texte contenant une nouvelle information ou pour coordonner deux (ou plusieurs) idées
C26	C32	Ajout d'un texte contenant la cause (la conséquence ou le but) d'une idée émise
C27	C33	Ajout d'un texte contenant une condition (une supposition ou une hypothèse)
C28	C34	Ajout d'un texte pour exprimer une opposition (une concession ou une restriction) par rapport à une idée émise
C29	C35	Ajout d'un texte pour expliquer (affirmer, illustrer ou résumer) une idée émise
C30	C36	Ajout d'un texte pour situer le moment ou la durée d'une idée émise
C37	C43	Suppression d'un texte qui contenait une nouvelle information
C38	C44	Suppression d'un texte qui contenait la cause (la conséquence ou le but) d'une idée émise
C39	C45	Suppression d'un texte qui contenait une condition (une supposition ou une hypothèse)
C40	C46	Suppression d'un texte dans lequel on avait exprimé une opposition (une concession ou une restriction) par rapport à une idée émise
C41	C47	Suppression d'un texte dans lequel on avait expliqué (affirmé, illustré ou résumé) une idée émise
C42	C48	Suppression d'un texte dans lequel on avait situé le moment ou la durée d'une idée émise

Tableau 3.3 – Catégories feuilles de la taxonomie des interventions

3.2.2 Organisation des interventions

La taxonomie des interventions que nous venons d'élaborer peut servir notamment à catégoriser des interventions des wikistes identifiées dans un corpus de logs.

Dans cette partie, nous allons voir comment ces interventions sont identifiées dans des historiques des modifications et comment elles sont organisées dans des structures adéquates avant cette étape de catégorisation.

Nous avons choisi de représenter les interventions des utilisateurs dans des séquences. En effet, une telle représentation permet de retrouver leur aspect temporel provenant de l'estampille de chaque version dans un historique de modification.

Une étape préliminaire à cette organisation des interventions est leur détection. Or, nous avons vu dans le chapitre précédent qu'une intervention est une séquence d'opérations de modification (ajout et suppression). Ainsi sa détection consiste à isoler la différence entre les deux paragraphes concernés (avec comme cas particuliers, ajout ou suppression d'un paragraphe entier).

Dans la suite, nous allons parler d'une technique d'isolation de différences entre deux fichiers textes (en l'occurrence deux versions successives d'un historique de modification correspondant à un article Wiki) proposée par *Hackel* [8] et dont nous nous sommes inspirés pour détecter les interventions. Par la suite nous exposerons l'algorithme *historySeq* qui permet de détecter différentes interventions dans un corpus d'historiques des modifications.

3.2.2.1 Détection de différence entre deux fichiers

D'après *Heckel* [8], il existe trois approches permettant de mettre en évidence la différence entre deux fichiers textes dont celle qu'il a proposée en 1978. Dans ce mémoire, nous nous sommes inspirés de cette technique dont nous présenterons le fonctionnement dans la suite de ce travail. Le choix de cette technique s'explique par le fait qu'elle donne de bons résultats sur les données étudiées (pages de Wikipedia) comparativement aux deux autres méthodes, notamment grâce à la détection d'un bloc entier de texte qui a été déplacé et à l'isolation juste des ajouts et des suppressions à l'intérieur de ce bloc au lieu de marquer tout le bloc comme un

changement. Ceci n'est pas possible avec les deux autres méthodes, en particulier, avec la technique classique de différence *LCS* (longest common subsequence). De plus c'est cette technique qui a été utilisée par *F. Viégas et al.* [21], un travail largement relié au notre et sa performance et les résultats obtenus sur les données de Wikipedia sont intéressants.

Cette approche consiste à trouver toutes les parties similaires entre deux fichiers et à considérer les parties restantes comme étant la différence recherchée. En prenant une ligne comme unité de base de comparaison (cette unité peut-être un caractère, un mot, un phrase ou un paragraphe), Paul Hackel fait deux observations :

1. *Une ligne qui apparaît une et une seule fois dans chacun des deux fichiers à comparer est la même ligne (inchangée mais éventuellement déplacée). Ceci permet d'identifier une grande partie de lignes qui sont ainsi exclues pour les calculs ultérieurs.*
2. *Si dans chacun des deux fichiers, immédiatement à la suite des paires de lignes identifiées précédemment, se trouvent des lignes identiques, ces dernières doivent être les mêmes. La répétition de cette observation permet d'identifier des blocs de lignes inchangées.*

De façon générale, cette technique s'applique à deux fichiers textes (ancien ou O et nouveau ou N) en supposant que O a été transformé en N et nécessite trois structures de données :

1. **Une table de symboles** (ou des différentes lignes). Un symbole est un identifiant d'une ligne (en l'occurrence le texte de la ligne) et sert de clé pour cette table. Chaque entrée de cette table contient deux compteurs *OC* et *NC* correspondant respectivement au nombre de copies de la ligne dans l'ancien et nouveau fichier. Ces compteurs valent 0, 1 ou "plusieurs". En plus de ces deux compteurs, l'entrée de cette table contient également un champ *OLNO* qui contient le numéro de la ligne dans l'ancien fichier. Cette valeur est utilisée uniquement dans le cas où OC=1.

2. **Un tableau** OA qui contient une seule entrée pour chaque ligne de l'ancien fichier. Il contient également soit un pointeur vers une entrée dans la table des symboles soit le numéro d'une ligne du nouveau fichier et un bit pour spécifier laquelle des deux valeurs. $OA[j]$ contient des informations relatives à la ligne j de O. Avec $j = 0$ et $j = o + 1$ correspondant aux deux lignes virtuelles au début et à la fin de O et o le nombre total des lignes de O.
3. **Un tableau** NAavec le même contenu que OA pour le nouveau fichier. Autrement-dit $NA[i]$ contient des informations relatives à la ligne i de N.

Ensuite, l'isolation de la différence entre les deux fichiers se fait par les six étapes suivantes :

1. La première étape consiste en quatre phases : premièrement, chaque ligne i du fichier N est lue, deuxièmement une entrée dans la table des symboles est créée pour chaque ligne i s'elle n'existe pas encore (vérification de la clé autrement-dit du texte de la ligne) puis la valeur de NC de cette ligne (symbole) est incrémentée. Afin $NA[i]$ est mis à jour de façon à pointer vers l'entrée dans la table des symboles correspondant à i.
2. La deuxième étape est identique à la première mis à part qu'elle concerne l'ancien fichier O (le tableau OA et la table des symboles : les nouveaux symboles, le compteur OC et $OLNO$).
3. La troisième étape s'inspire de la première observation vue plus haut et traite uniquement les lignes dont $NC = OC = 1$. Ces lignes étant considérées comme inchangées, pour chacune d'entre elles, on remplace le pointeur vers la table des symboles, dans NA et OA, par le numéro de la ligne dans l'autre fichier ainsi que le bit qui désigne l'information contenue. Par exemple, si la ligne en question est $NA[i]$, il faut regarder l'entrée correspondant à $NA[i]$ dans la table des symboles et modifier $NA[i]$ pour lui donner la valeur de $OLNO$ et modifier $OA[OLNO]$ en i. Deux lignes virtuelles, l'une au début et l'autre à la fin de chaque fichier, sont également considérées et identifiées comme étant inchangées durant cette phase.

4. La quatrième étape s'inspire de la deuxième observation et permet de retrouver les autres lignes inchangées mais qui apparaissent plusieurs fois dans O ou N. Pour cela, on considère chaque ligne de NA dans l'**ordre ascendant** : si $NA[i]$ pointe vers $OA[j]$ et que $NA[i+1]$ et $OA[j+1]$ contiennent des pointeurs vers la même entrée dans la table des symboles alors $OA[j+1]$ est modifié en $i+1$ et $NA[i+1]$ et modifié en $j+1$. Notons que si $NA[1]$ et $OA[1]$ contiennent des pointeurs vers la même entrée dans la table des symboles alors ces deux lignes seront reconnues comme correspondantes car $NA[0]$ pointe vers $OA[0]$.
5. La cinquième étape est identique à la précédente mis à part que cette fois-ci le tableau NA est parcouru en **ordre descendant**.
6. A la fin de ces précédentes cinq étapes, le tableau NA contient tout ce qu'il faut pour lister les différences entre les deux fichiers. Durant la dernière étape, on parcourt NA, si $NA[i]$ pointe vers une entrée dans la table des symboles alors il s'agit d'une ligne ajoutée. Sinon s'il pointe vers $OA[j]$ et que $NA[i+1]$ ne pointe pas vers $OA[j+1]$, alors la ligne i est à la limite d'une suppression ou d'un bloc de lignes déplacées. Les lignes supprimées peuvent être retrouvées en parcourant OA. Ainsi de suite jusqu'à ce que tout le tableau NA soit parcouru.

3.2.2.2 Technique *historySeq*

Pour ce mémoire, il était important de tenir compte de la taille des fichiers à comparer à savoir les versions d'articles de Wikipedia qui sont dans certains cas très larges. La technique de Paul Heckel introduite précédemment présente un avantage à ce niveau par le fait qu'elle peut être utilisée pour comparer des fichiers très longs. En effet, parmi les trois structures utilisées, seule la table des symboles qui est parcourue de temps en temps, a une taille qui augmente suivant la taille des deux fichiers à comparer. Pour remédier à cela, la taille de chaque entrée de cette table peut être diminuée en combinant les fonctions des deux compteurs NC et OC en un seul champ et en éliminant le champs $OLNO$.

Cependant, cette même technique est susceptible de détecter de fausses différences. En effet, supposons un bloc de lignes inchangées et qui ne sont pas uniques (elles apparaissent plusieurs fois dans N ou O). Si les lignes qui précèdent et suivent ce bloc ont changé, elles seront (correctement) détectées comme une différence, cependant le bloc (de lignes inchangées) qui se trouve entre elles sera aussi (faussement) marqué comme une différence. En effet, le bloc de lignes n'a pas pu être détecté durant la quatrième phase ou la cinquième. Bien que cette situation ne soit pas trop grave, elle peut être évitée en utilisant une unité de base de comparaison différente (par exemple un bloc de trois lignes) ou une hiérarchie de plusieurs unités.

Dans ce travail, nous avons considéré deux unités de comparaison : un paragraphe et un mot. En effet, durant nos observations des historiques des modifications avec l'outil *historyDiff*, nous avons constaté que les interventions des utilisateurs d'une version à la suivante d'un article se concentrent sur quelques paragraphes et le reste du texte demeure inchangé. Ainsi en prenant le paragraphe comme unité de comparaison nous calculons de façon assez rapide la différence entre deux versions successives d'un article.

En prenant un paragraphe comme unité de comparaison et en procédant suivant la technique de Paul Heckel, nous pouvons détecter : des nouveaux paragraphes, des paragraphes supprimés et des paragraphes déplacés. Cependant avec cette méthodologie, un paragraphe modifié n'est pas détecté en tant que tel mais comme une combinaison d'un paragraphe supprimé et un nouveau paragraphe ajouté. Pour remédier à cette situation, nous détectons un paragraphe modifié et calculons la différence entre les deux paragraphes (celui de l'ancienne version et son correspondant dans la nouvelle version) en prenant cette fois-ci un mot comme unité de comparaison.

La figure 3.7 à la page 62 illustre le fonctionnement de notre technique. Dans chaque version, nous avons des paragraphes contenant un seul mot (par exemple,

la première version est constituée de 3 paragraphes : Banane, Kiwi et Pommes). De la première version à la deuxième, les deux premiers paragraphes sont restés inchangés, par contre le troisième a été modifié (la différence calculée au niveau du mot pour ce paragraphe est : suppression de "pommes" et insertion de "pomme"). Il y a eu également une insertion d'un nouveau paragraphe "Orange".

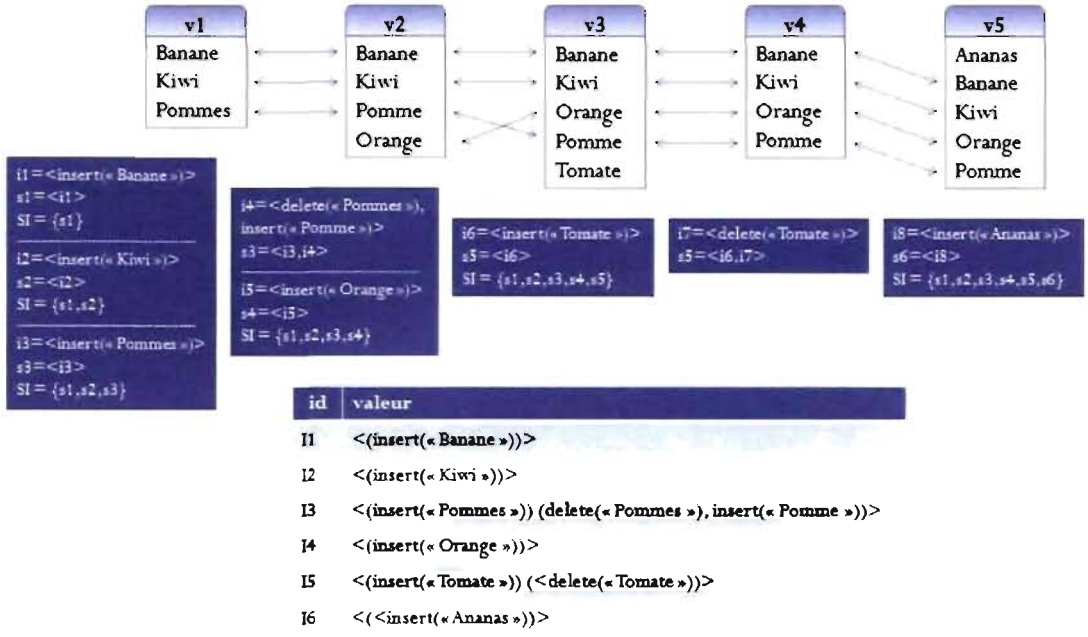

Figure 3.7 – Détection et organisation des interventions

La technique *historySeq* proposée, consiste donc à la détection et organisation des interventions d'un historique des modifications d'une page Wiki. Une variante de celle-ci permet de procéder à la détection et organisation des interventions issues de plusieurs historiques des modifications.

Algorithme 1 Algorithme *historySeq*

Entrées: $H = \{h_1, h_2, \ldots, h_n\}$ ensemble de n historiques des modifications
Sorties: $DI = \{I_1^1, I_2^1, \ldots, I_{m_1}^1, I_1^2, \ldots, I_{m_2}^2, \ldots, I_1^n, \ldots, I_{m_n}^n\}$ base de séquences d'interventions. I_k^l, la séquence d'interventions du paragraphe k de l'historique h_l

```
pour chaque historique h_l, 1 ≤ l ≤ n faire
  pour chaque version v_s^l, 1 ≤ s ≤ |h_l| (le nombre de versions dans h_l) faire
    si s = 1 (1^ére version de h_l) alors
      pour chaque paragraphe P_{m_l}^l faire
        Créer une séquence d'interventions I_{m_l}^l
        Créer une intervention ι :< insert(t_{ml}^l) >
        Ajouter ι dans I_{m_l}^l
        Ajouter I_{ml}^l dans DI
      fin pour
    sinon
      Apparier les textes des paragraphes de v_s^l et ceux de v_{s-1}^l
      Détecter la différence ou mettre en évidence les interventions
    finsi
  fin pour
fin pour
```

Les instructions à l'intérieur de la boucle "si" forment ce que nous pouvons appeler la phase d'**initialisation** tandis que les deux instructions situées dans la boucle "sinon" forment la **méthode**. La détection de la différence consiste en l'isolation des unités non-appariées qui sont alors considérées comme étant la différence entre les deux versions. Trois détections possibles :

- `Paragraphe ajouté (`P_a`) : Même traitement que pour la phase d'initialisation.`
- `Paragraphe supprimé (`P_s`) : Créer et ajouter dans la séquence d'interventions` I_s`, une intervention` $\iota :< delete(t_s) >$.
- `Paragraphe modifié (`P_m`) : Appariement des` *mots* `de ce paragraphe dans` v_s^l `et ceux dans` v_{s-1}^l `puis détection de différence. Deux cas possibles :`
  - `mots ajoutés : Créer et ajouter dans dans la séquence d'interventions` I_m`, une intervention` $\iota :< insert(t) >$;
  - `mots supprimés : Créer et ajouter dans dans la séquence d'interventions` I_m`, une intervention` $\iota :< delete(t) >$.

3.2.3 Catégorisation des interventions

Une fois les interventions repérées et organisées dans des séquences, nous procédons à leur catégorisation. Cette étape a pour effet d'uniformiser les interventions dans les séquences (en leur donnant des étiquettes contenant une description plus générale autrement-dit une catégorie des interventions) afin de pouvoir extraire des motifs fréquents. La procédure que nous avons nommée *historyCat*, utilise donc la taxonomie des interventions afin de classifier chacune des interventions se trouvant dans les séquences.

3.2.3.1 Technique *historyCat*

La technique *historyCat* procède de la façon suivante :

Algorithme 2 Algorithme *historyCat*

Entrées: $DI = \{I_1, I_2, \ldots, I_n\}$ (une base de séquences d'interventions) et τ (Taxonomie des interventions)

Sorties: $DS = \{S_1, S_2, \ldots, S_n\}$, une base de séquences de données. Où S_k, avec k de 1 à n, est la séquence de données correspondant à la séquence d'interventions I_k

pour chaque séquence d'interventions $I_k, 1 \leq k \leq n$ **faire**
  Créer une séquence de données S_k vide
  pour chaque intervention $\imath_l^k, 1 \leq l \leq |I_k|$ (nombre d'interventions) **faire**
    Créer un categorieset s_l^k
    pour chaque catégorie d'interventions feuille $c_i \in \tau$ **faire**
      si description de c_i = description d'une des opérations de modification composant l'intervention $\imath_l^k$ **alors**
        si $c_i \notin s_l^k$ **alors**
          Ajouter c_i dans s_l^k
        finsi
      finsi
    fin pour
    Ajouter s_l^k dans S_k
  fin pour
  Ajouter S_k dans DS
fin pour

3.2.4 Extraction des motifs

Une fois le prétraitement et la transformation des données effectués, nous procédons à l'extraction des motifs de collaboration fréquents. De façon générale, des motifs de collaboration (ou des séquences de catégories d'interventions) de différente taille sont générés en utilisant la taxonomie des interventions et leurs fréquences sont calculées par rapport aux séquences de données obtenues lors de l'étape précédente. Les motifs de collaboration fréquents sont alors retenus. Le paragraphe suivant explique le fonctionnement de ce procédé que nous avons appelé *historyPattern*.

3.2.4.1 Technique *historyPattern*

L'extraction de motifs de collaboration se fait comme suit :

Algorithme 3 Algorithme *historyPattern*

Entrées: $DS = \{sc_1, sc_2, \ldots, sc_n\}$ (base de séquences de données), τ (Taxonomie des opérations de modification) et σ le support minimum

Sorties: $M_\sigma = \{m_1, m_2, \ldots, m_l\}$ ensemble de l motifs de collaboration fréquents

```
Générer et stocker des motifs candidats de rang 1 dans MC_1
pour chaque motif candidat mc_j^1 ∈ MC_1 faire
  si supp(mc_j^1) > σ alors
    Ajouter mc_j^1 dans M_σ1
  finsi
fin pour
k ← 2
tantque M_σk−1 n'est pas vide faire
  Générer et stocker des motifs candidats de rang k dans MC_k
  pour chaque candidat mc_j^k faire
    si supp(mc_j^k) > σ alors
      Ajouter mc_j^k dans M_σk
    finsi
  fin pour
  k = k + 1
fin tantque
```

Les instructions avant $k \leftarrow 2$ constituent la phase d'initialisation alors que l'instruction $k \leftarrow 2$ ainsi que toutes les instructions après constituent la méthode.

La phase de "Génération des motifs candidats" est effectuée de la façon suivante :

- génération des candidats de rang 1 (MC_1) : consiste à prendre des séquences formées d'un *categorieset* singleton : $< \{c\} >$ tel que $d(c_i) = 1$;
- génération des candidats de rang supérieur à 1 : consiste à appliquer trois opérations canoniques à un ensemble de motifs fréquents de façon à former de nouveaux motifs dont le rang a été augmenté de 1 autrement-dit $rang(motifrsultant) = rang(motifd'origine) + 1$. Ces opérations sont :
  - $add(\{c\})$: ajout d'un *categorieset* contenant une catégorie de degré 1.
  - $add(c)$: ajout d'une catégorie d'interventions dont le degré = 1, au dernier catégorieset (c différente des autres catégories de ce catégorieset).
  - $spec(c)$: spécialisation (d'un niveau) de la dernière catégorie d'interventions du dernier catégorieset (remplacement par le prédécesseur immédiat pour $\leq_T$, c doit être une catégorie intermédiaire).

Une fois les motifs candidats générés, nous calculons leur support, ce qui renvient à calculer toutes les séquences de données que couvre le motif candidat. Ensuite ce support est comparé au support fixé par l'utilisateur. Les motifs candidats dont le support dépasse le support minimum sont considérés comme fréquents.

3.2.5 Visualisation des motifs de collaboration fréquents

Afin d'interpréter les motifs de collaboration fréquents extraits, l'utilisateur a besoin de les visualiser. Comme ces motifs de collaboration fréquents sont sous forme de séquences, l'utilisateur doit disposer d'un outil de visualisation qui prend en compte des séquences. De plus, l'utilisateur doit pouvoir distinguer les différents aspects des interventions (ajout ou suppression, long texte ou texte court, la relation sémantique exprimée). Ceci peut être géré en associant par exemple une couleur spécifique à chaque aspect d'une intervention. Ainsi chaque catégorie d'une intervention peut être représentée par une boîte d'une couleur qui lui est propre et une série de plusieurs boites représenterait alors une séquence de catégories d'interventions et en particulier un motif de collaboration fréquent.

CHAPITRE 4

L'OUTIL *HISTORYMINER*

Dans cette section nous allons parler de l'outil *historyMiner* qui est une mise en pratique de la méthodologie proposée dans ce mémoire. Nous commencerons par son implémentation et présenterons ensuite ses fonctionnalités.

4.1 Implémentation

L'outil *historyMiner* a été implémenté en Java. Le choix de ce langage s'explique par notre expertise dans ce domaine et aussi par le fait que la méthodologie à implémenter n'exigeait aucun langage en particulier. Ce choix permet également une grande flexibilité au niveau de l'utilisation de l'outil car un fichier exécutable "jar" est généré et l'utilisateur n'a besoin que de la machine virtuelle Java pour l'exécuter (ce programme est accessible gratuitement sur Internet dans le cas où il n'est pas déjà installé sur l'ordinateur). Un autre avantage de ce langage est la facile d'intégration d'autres fonctionnalités. Ceci nous a permit notamment d'utiliser le package JGraph [1] pour implémenter la visualisation.

4.2 Les différentes fonctionnalités de cet outil

L'outil *historyMiner* présente 2 fonctionnalités principales : la visualisation des interventions d'un historique des modifications (d'un article ou page d'un site Wiki) et l'extraction et la visualisation de motifs de collaboration fréquents issus d'un ensemble d'historique des modifications. La version actuelle de l'outil supporte uniquement le format XML des historiques des modifications proposées par le moteur de Wiki, Mediawiki uniquement.

[1] http ://www.jgraph.com/

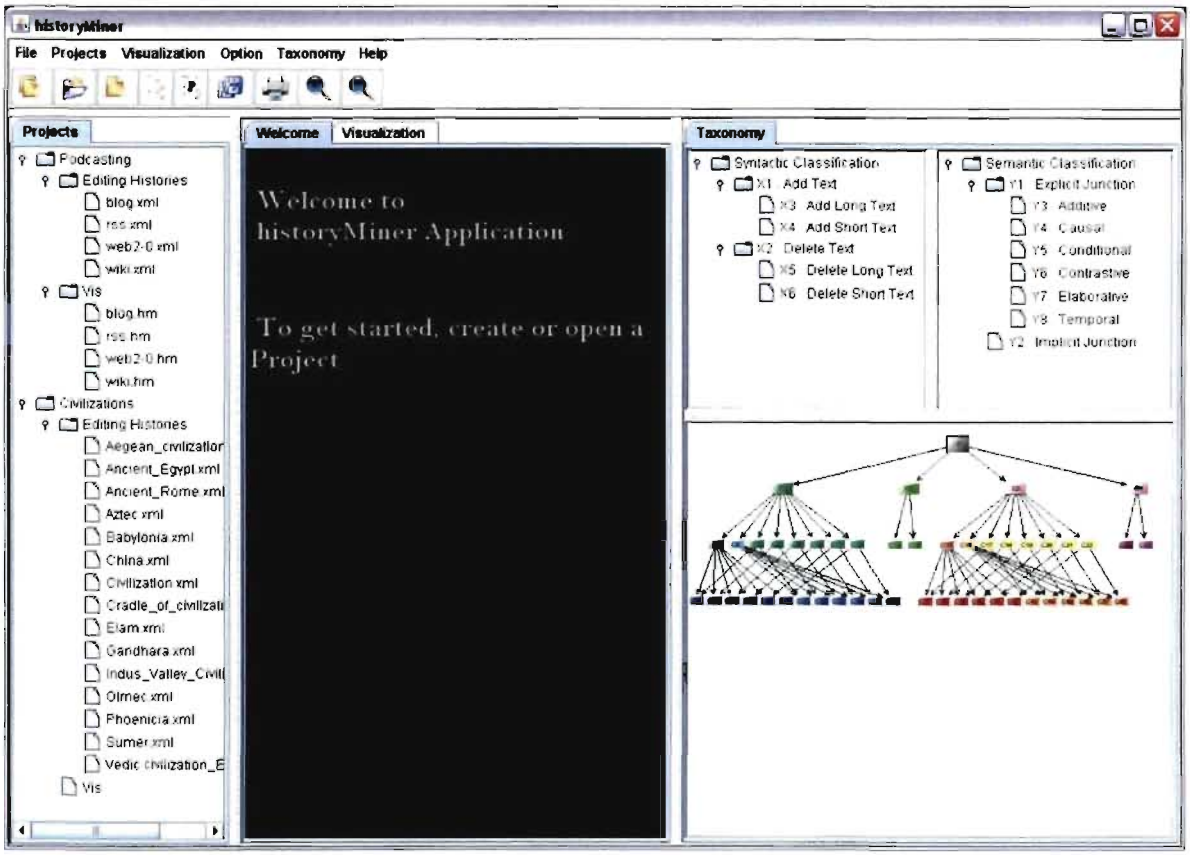

Figure 4.1 – Page d'accueil - *historyMiner*

Comme nous pouvons le constater sur la figure 4.1 à la page 68, l'outil *historyMiner* est divisé en deux parties. La partie supérieure composée d'un menu et d'une barre d'outils puis la partie inférieure qui est divisée en trois cadres :

- un cadre d'exploration des projets nommé "Projects" ;
- un cadre pour la visualisation (avec deux onglets : "Welcome" pour l'affichage d'un message de bienvenue et "Visualization" pour la visualisation des motifs de collaboration fréquents ou d'un historique des modifications) ;
- un cadre "Taxonomy" pour la consultation de la taxonomie des interventions.

Un projet dans *historyMiner* est un répertoire comportant deux sous-répertoires :

- ***Editing History*** : un répertoire comportant des historiques des modifications des pages d'un Wiki. Chaque historique des modifications est un fichier XML (identique à celui ce la figure 3.2 à la page 47). Dans l'exemple de la figure 4.1 à la page 68, nous avons deux projets : *Podcasting* avec 4 historiques des modifications (blog.xml, rss.xml, web2-0.xml et wiki.xml) et *Civilizations*.
- ***Vis*** : le répertoire dans lequel *historyMiner* sauvegarde toutes les visualisations correspondantes au projet. Par exemple, "blog.hm" est la visualisation des interventions identifiées dans l'historique des modifications "blog.xml" et "project1.hm" est la visualisation des motifs de collaboration fréquents extraits dans les quatres articles faisant partie de ce projet.

4.2.1 Visualisation des interventions d'un historique des modifications

Cette fonctionnalité permet à l'utilisateur de visualiser les interventions ayant eu lieu sur un article, c'est-à-dire les interventions issues de son historique des modifications.

Chaque intervention est représentée sous forme d'un groupe (cluster) de cases correspondant chacune à une catégorie d'interventions à laquelle appartient l'intervention. Une colonne de clusters correspond aux interventions d'un wikiste sur une version spécifique de l'article tandis qu'une ligne de clusters représente les interventions de différents utilisateurs sur un paragraphe spécifique de l'article.

Cette fonctionnalité est en quelque sorte une visualisation avancée d'un historique des modifications telle que présentée dans l'outil history flow. On peut d'ailleurs le constater en regardant l'allure du graphe obtenu par les deux logiciels sur la figure 4.2 à la page 70.

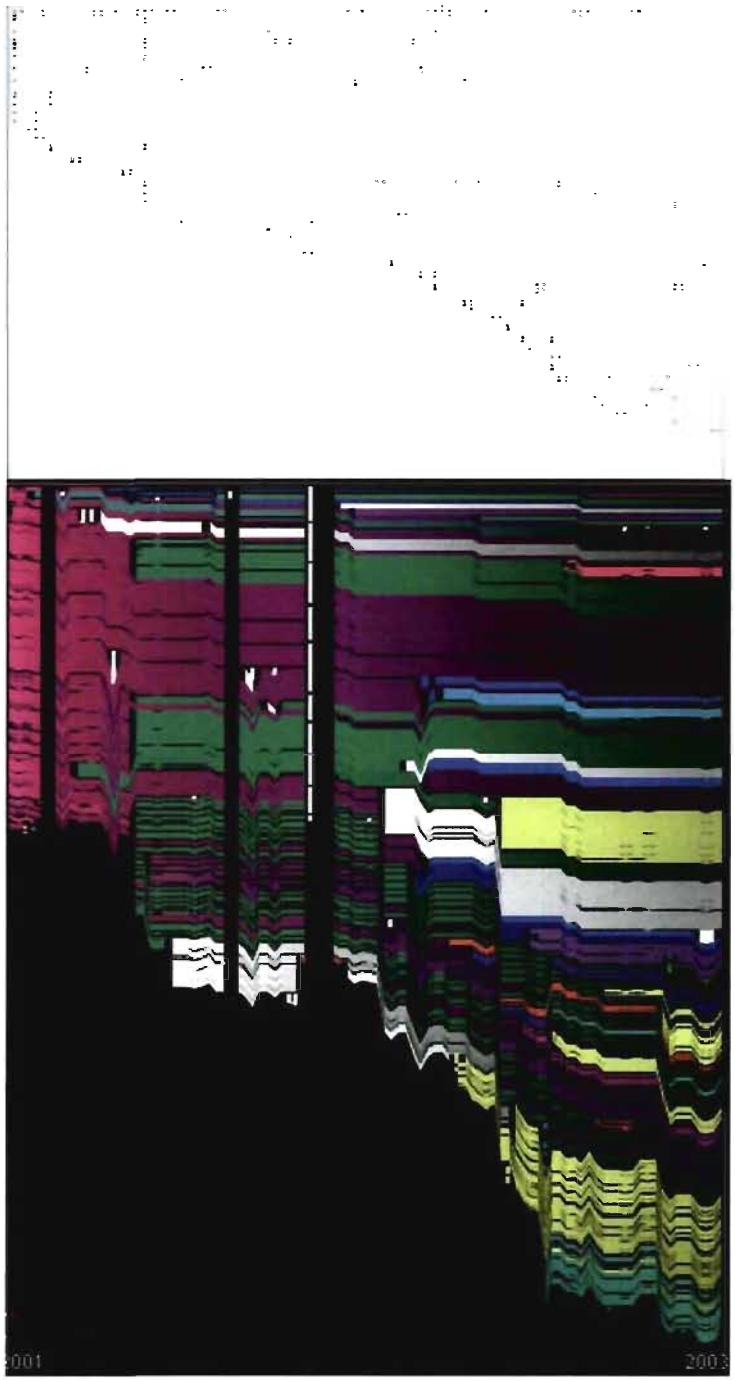

Figure 4.2 – Visualisation de l'article "Wiki" avec *historyMiner* et History Flow

Dans ces deux visualisations de l'article "Wiki" de la version anglaise de Wikipedia, nous pouvons constater que le contenu des premiers paragraphes de l'article a persisté (absence de clusters dans *historyMiner* et des lignes continues dans history flow), l'article a plutôt subi des ajouts de nouveaux paragraphes au fur du temps. Aux informations données par history flow concernant ces interventions (l'auteur, le texte, etc.), *historyMiner* rajoute la description générale des interventions suivant l'opération de modification et le lien sémantique avec les interventions précédentes.

4.2.2 Extraction et visualisation des motifs de collaboration fréquents

La fonctionnalité la plus importante de l'outil *historyMiner* est l'extraction des motifs de collaboration fréquents et la visualisation de ces derniers afin de pouvoir les interpréter. Les motifs de collaboration fréquents sont, comme nous l'avons mentionné, des séquences d'ensemble de catégories d'interventions et sont générés suivant un rang croissant. La visualisation proposée dans la version actuelle de *historyMiner* suit ces deux points. Ainsi, comme nous pouvons le constater sur la figure 4.3 à la page 72, chaque ligne représente une séquence de *catégoriesets* ou motifs de collaboration.

Plus exactement, chaque ligne représente un comportement fréquent de plusieurs paragraphes (suivant le support minimum fixé par l'utilisateur) des pages d'un site Wiki en l'occurrence, des articles de la version anglaise de Wikipedia. Si le support minimum est "20", la $1^{ère}$ ligne C13 signifie que 20% des paragraphes analysés ont subi une intervention d'ajout d'un grand texte sans exprimer la relation sémantique avec le texte des interentions précédentes. Remarquons que l'utilisateur dispose de la taxonomie des interventions sur le coté droit afin de voir la description de chaque catégorie faisant partie des motifs de collaboration fréquents.

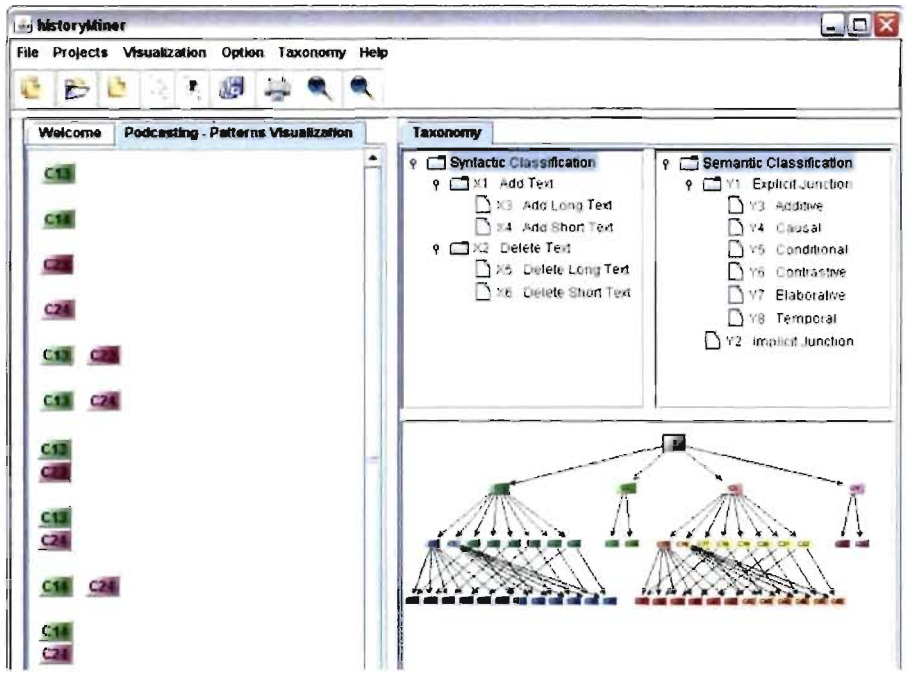

Figure 4.3 – Visualisation des motifs de collaboration fréquents - *historyMiner*

4.2.3 Caractéristiques supplémentaires

En plus de ces deux fonctionnalités, l'utilisateur a la possibilité de :

- faire un "zoom" avant-en arrière de la visualisation en cours ;
- imprimer la visualisation en cours ;
- faire un "zoom" sur la taxonomie des interventions. De plus un clic sur chaque catégorie d'interventions met en évidence les deux valeurs de la classification syntaxique et sémantique correspondantes ;
- sauvegarder une visualisation (d'un historique des modifications ou des motifs de collaboration fréquents). Ceci permet à l'utilisateur de sauver du temps en lui évitant notamment l'extraction de motifs de collaboration fréquents pour le même projet, à chaque fois qu'il veut les analyser ;
- spécifier la valeur du support minimum ;
- consulter une aide complète sur le fonctionnement du logiciel.

CHAPITRE 5

EXPÉRIMENTATIONS ET VALIDATION

Dans cette section nous allons parler des expérimentations effectuées avec l'outil *historyMiner* sur un corpus d'articles venant de Wikipedia.

5.1 Choix du corpus

Afin de valider notre méthodologie, nous avons effectué quelques expérimentations avec l'outil *historyMiner*. La première étape était le choix du corpus. Nous avons considéré quelques caractéristiques des articles à savoir : la taille des articles (version finale), le nombre de versions de chaque article, la période de développement d'un article (différence entre l'estampille ou date de création de la première version et celle de la dernière version) et l'argumentation dans le contenu.

En effet, pour pouvoir extraire des motifs de collaboration fréquents intéressants, il fallait trouver des articles avec un contenu riche (de grande taille) et argumenté (avec des connecteurs pour marquer les relations sémantiques entre les idées). Pour évaluer l'argumentation, nous avons fait une étude empirique visant à évaluer l'utilisation des connecteurs exclusifs en calculant leur fréquence d'apparition sur deux corpus.

Le premier échantillon d'articles nous a été proposé par nos collaborateurs du Département de Communication et était composé par 16 articles autour du terme "podcasting", tableau 5.1 à la page 74.

Article	Nombre de versions	Estampille première version	Estampille dernière version
1. Adam Curry	555	17 :11, 8 May 2004	06 :57, 26 August 2006
2. Aggregator	265	02 :37, 1 August 2005	18 :29, 25 August 2006
3. Atom (standard)	322	23 :12, 6 March 2004	00 :37, 22 August 2006
4. Copy Protection	312	09 :38, 22 April 2003	21 :07, 28 August 2006
5. Copyright	1199	06 :53, 30 September 2001	18 :02, 30 August 2006
6. Creative Commons	429	12 :54, 17 May 2002	23 :39, 27 August 2006
7. Digital Rights Management	1219	03 :24, 22 September 2002	23 :12, 30 August 2006
8. MP3	1334	17 :09, 30 September 2001	14 :19, 31 August 2006
9. Online music store	356	03 :46, 18 October 2003	13 :19, 29 August 2006
10. Podsafe	79	19 :32, 21 July 2005	01 :43, 22 August 2006
11. Public domain	714	00 :35, 5 November 2001	08 :39, 31 August 2006
12. Recording Industry Association of America	734	02 :31, 20 February 2002	15 :33, 30 August 2006
13. Ripping	132	23 :07, 6 March 2004	14 :51, 22 August 2006
14. Royalties	34	06 :14, 27 August 2003	15 :05, 27 July 2006
15. RSS (file format)	1378	23 :32, 27 September 2002	13 :18, 31 August 2006
16. Windows Media Audio	213	19 :53, 10 January 2002	08 :41, 31 August 2006

Tableau 5.1 – Corpus 1 : articles de Wikipedia autour du terme "Podcasting"

Pour obtenir les différentes versions de chacun des articles, nous avons d'abord essayé de télécharger les "dumps" proposés par Wikipedia[1]. Malheureusement nous nous sommes vite retrouvés confrontés au problème de la grande taille de ces dumps (il était impossible d'obtenir un seul article à la fois, il fallait télécharger une version complète de Wikipedia ce qui veut dire dans notre cas, Wikipedia en anglais complet). Les tentatives de téléchargement ont amené à des dumps incomplets ou illisibles. Nous avons alors programmé un "robot" qui cherche version par version chacun des articles et les enregistre en local. Le tableau 5.1 à la page 74 contient des articles téléchargés le 1 septembre 2006.

Les résultats sur l'utilisation des connecteurs exclusifs dans les 16 articles (voir le

[1]Wikipedia dumps : http ://download.wikimedia.org/backup-index.html

tableau de l'annexe I) montrent notamment que seulement un certain nombre des connecteurs considérés est fréquemment utilisé. Parmi ceux-ci nous retrouvons : *although, because, but, even, first, for, if, now, rather, so, still, though, thus, yet.* A savoir que d'autres connecteurs comme *and, as, or, when, where, while* apparaissaient régulièrement dans les 16 articles mais n'ont pas été considérés car ils font partie des connecteurs partagés entre plusieurs relations et donc ils ne nous permettent pas de détecter de façon automatique les relations exprimées.

La conclusion de cette première expérimentation est que les relations sémantiques sont plutôt implicites qu'explicites. Nous avons donc décidé de faire une autre étude empirique, cette fois-ci avec un corpus plus riche et dans un autre domaine.

Après quelques consultations des articles de Wikipédia, nous avons opté pour le domaine de l'histoire et nous avons considéré les articles de la catégorie "Civilizations" qui dans la version anglaise de Wikipedia, sont assez dévéloppés. En effet, la catégorie "Civilizations" comptait 35 pages au 15 avril 2008 ainsi que 38 sous-catégories (voir l'images en annexe III) qui, à leur tour, contenaient **1 143 pages** différentes.

Pour nos expérimentations, nous nous sommes limités aux 115 pages les plus volumineux (le nombre de caractères pour la dernière version était supérieur à 20 000 sans considérer les parties à la fin de la page comme "See also", "References", etc.). En utilisant la fonctionnalité d'exportation des pages[2], nous avons pu télécharger les 100 premières versions de chaque page. Dans la figure 5.1 à la page 76, nous avons les noms de ces pages et le tableau de l'annexe II contient les résultats de notre deuxième étude sur l'évaluation de l'utilisation des connecteurs exclusifs dans ce deuxième corpus.

[2]http ://en.wikipedia.org/wiki/Special :Export. Cette fonctionnalité n'existait pas lors de l'exportation du 1er corpus

Adana	Cuneiform script	Mycenaean Greece
Aegean civilization	Damascus	Names of China
Age of Enlightenment	Dead Sea scrolls	Neo-Assyrian Empire
Air Nomads	Decline of the Roman Empire	Nochiya Tribe
Akkad	Dorian invasion	Norte Chico civilization
Amarna letters	Earth Kingdom	Olmec
Ancient Egypt	Elam	Pelasgians
Ancient Greece	Etruscan civilization	Phaistos Disc
Ancient Greek	Fire Nation	Phoenicia
Ancient Rome	Gandhara	Pop music
Ancient Rome and wine	Great Books of the Western World	Postal history of Palestine
Antioch	Greece	Postmodernity
Antioch, Pisidia	History of ancient Egypt	Pottery of ancient Greece
Aramaic language	History of pottery in the Southern Levant	Prehistoric Cyprus
Archaeology and the Book of Mormon	History of saffron	Pre-Roman history of ancient Israel and Judah
Ayurveda	History of Sumer	Roman historiography
Aztec	History of Western civilization	Roman technology
Aztec society	Humans	Saffron
Babylonia	Inca Empire	Samothrace temple complex
Babylonian astronomy	Indus River	Southern Maya area
Babylonian law	Indus Valley Civilization	Sumer
Balaam	Iraq	Sumerian king list
Battle of Kadesh	Kingdom of Kongo	Sumerian language
Beirut	Kirkuk	The Culture
Carthage	Linear B	The Decline of the West
Chania	Macedon	The History of the Decline and Fall of the Roman Empire
China	Macedonia (region)	The Story of Civilization
Christianity	Maya civilization	Time Lord
Chronology of the ancient Near East	Mesopotamia	Timeline of ancient Rome
Civilization	Military history of the Neo-Assyrian Empire	Tripoli, Lebanon
Clash of Civilizations	Minoan chronology	Troy
Classic Maya collapse	Minoan civilization	Urartu
Classical Greece	Minoan eruption	Varna
Cradle of civilization	Minoan pottery	Vedic civilization/EB 1911
Cretan War	Misthi	Vulgar Latin
Crucifixion	Mitanni	Water Tribe
Cultural and historical background of Jesus	Morlock	Western culture
Culture by region	Mosul	
Culture of ancient Rome	Muqaddimah	

Figure 5.1 – Corpus 2 : articles de Wikipedia dans la catégorie "Civilizations"

Les résultats de cette deuxième expérimentation nous ont révélés trois choses dont la confirmation de nos deux constatations lors du premier test : un ensemble restreint de connecteurs est plus utilisé par les wikistes. En effet, nous avons constaté que les connecteurs les plus utilisés dans les deux cas étaient les mêmes ; les relations sémantiques sont plus implicites qu'explicites (pas beaucoup de connecteurs exclusifs ce qui de plus pourrait affecter la méthodologie proposée car elle se base entre autres sur la présence des connecteurs pour catégoriser les interventions). Cependant, un nouveau fait qui se dégage de ce deuxième test est que le nombre de connecteurs dépend du corpus, en effet nous avons constaté une présence accrue

des marqueurs de sens dans le deuxième corpus.

5.2 Motifs de collaboration fréquents

Pour la suite de nos expérimentations, nous avons donc retenu le deuxième corpus qui était riche en terme de nombre de pages et de connecteurs. Afin d'extraire des motifs de collaboration fréquents en utilisant l'outil *historyMiner*, les catégories (de pages Wikipédia) dans lesquelles appartenaient les 115 pages ont constitué les "Projects" dans *historyMiner*.

Par exemple la catégorie "Civilizations" de Wikipedia contient 35 pages mais dans le projet "Civilizations" nous n'avons considéré que les 15 pages les plus volumineuses de cette catégorie, à savoir : *Aegean civilization*, *Ancient Egypt*, *Ancient Rome*, *Aztec*, *Babylonia*, *China*, *Civilization*, *Cradle of civilization*, *Elam*, *Gandhara*, *Indus Valley Civilization*, *Olmec*, *Phoenicia*, *Sumer*, *Vedic civilization/EB 1911*.

L'ensemble des motifs de collaboration fréquents obtenus est dominé par les catégories "C13", "C14","C23" et "C24" qui correspondent à l'ajout/suppression d'un texte court/long et sans marqueur pour exprimer la relation sémantique. Autrement-dit dans une très grande partie des interventions analysées, les utilisateurs n'avaient pas utilisé des connecteurs au début de leurs interventions. Le tableau ci-dessous représente quelques motifs obtenus avec un support minimum variant entre 20% et 60% (20% des paragraphes ont eu un tel comportement tout au long de l'évolution de l'article) et impliquant les catégories plus spécifiques (catégories feuilles de la taxonomie).

Id	**Motif**
1.	C13 : <{(Add long text, Implicit)}>
2.	C14 : <{(Add short text, Implicit)}>
3.	C23 : <{(Delete long text, Implicit)}>
4.	C24 : <{(Delete short text, Implicit)}>
5.	C13 C24 : <{(Add long text, Implicit)},{(Delete short text, Implicit)}>
6.	C14 C24 : <{(Add short text, Implicit)},{(Delete short text, Implicit)}>
7.	C13 C14C24 : <{(Add long text, Implicit)},{(Add short text, Implicit),(Delete short text, Implicit)}>
8.	C14 C14C24 : <{(Add short text, Implicit)},{(Add short text, Implicit),(Delete short text, Implicit)}>

Tableau 5.2 – Motifs de collaboration fréquents

Les deux premiers motifs représentent en grande majorité l'ajout de nouveaux paragraphes dans les premières versions des articles. Ces nouveaux paragraphes subissent quelques temps après soit une suppression d'un long texte, soit une suppression d'un petit texte ou encore les deux opérations en même temps. Ceci est marqué par la présence des motifs 5, 6, 7 et 8.

Ces constats nous poussent à croire que les premières versions des articles sur Wikipedia ne sont qu'une liste de plusieurs idées qui vont, une à une, être développées au fur du temps par d'autres utilisateurs. Le lien entre les idées (paragraphes) reste implicite ou n'est pas fréquemment marqué. Plus généralement, les utilisateurs manifestent rarement le sens de leur intervention par rapport au texte existant.

CHAPITRE 6

CONCLUSION

Le travail effectué durant ce mémoire peut être conclu en deux parties. Dans la première section, nous rappelons les principales contributions à la recherche de mécanismes de collaboration au sein des sites Web Wiki et revenons sur les discussions réalisées au cours du mémoire. Nous présentons au courant de la deuxième section les perspectives associés à ce travail.

6.1 Contributions de ce travail

Au cours de ce mémoire, nous avons abordé la problématique de la recherche des mécanismes de collaboration au sein des sites Web Wikis.

Après avoir décrit au chapitre 2, les mécanismes de collaboration, nous avons montré qu'une façon de les identifier est de rechercher leurs manifestations. Ces dernières sont par la suite interprétées par un expert (en l'occurrence un chercheur en communication) afin de déterminer les mécanismes de collaboration recherchés.

Nous avons également mentionné que les manifestations des scénarios collaboratifs recherchés sont des structures typiques d'interventions des utilisateurs et se trouvent dans les historiques des modifications. De plus, les interventions des utilisateurs sont spécifiques au texte de chaque page. Afin d'identifier des mécanismes de collaboration à l'échelle d'un site Wiki, au lieu d'une page, nous avons introduit la notion de catégorie d'interventions et nous avons ainsi posé la problématique de la recherche de séquences de catégories d'interventions typiques ou motifs de collaboration fréquents.

Au cours de ce chapitre, nous avons ainsi examiné les différents travaux autour des mécanismes de collaboration sur les Wikis, et plus généralement, les travaux autour de la qualité des contenus sur les sites Wikis. Nous en avons également profité pour introduire la notion de cohérence textuelle dont nous nous sommes inspirée pour créer les catégories d'interventions. Enfin, nous avons fait un rappel sur les notions de base de la fouille de données, nous sommes aussi revenus sur les travaux autour des règles d'association ainsi que sur les motifs séquentiels, dont s'est inspirée notre problématique d'extraction de motifs de collaboration fréquents.

Dans le Chapitre 3, nous avons présenté l'approche proposée. Après avoir montré l'architecture qui s'inspire des 5 étapes de KDD ou recherche des connaissances dans les bases de données, nous avons vu en détail chacune des 5 tâches principales de cette architecture.

Tout d'abord une taxonomie des interventions sur un outil de collaboration tel qu'un site Wiki ainsi que son élaboration sont proposées. Cette taxonomie présente les différentes catégories d'interventions des participants lors de la rédaction collaborative des contenus. Pour déterminer la catégorie d'interventions, deux aspects d'une intervention sont considérés : d'un côté, la syntaxique (qui relève des règles d'écriture sur l'outil. Exemple : insertion ou suppression d'un texte dans un document) de l'autre côté, la sémantique ou le rôle d'une intervention (qui relève plus des règles linguistiques d'écriture. Exemple : l'opposition, l'élaboration, etc.).

Cette taxonomie est ensuite utilisée pour catégoriser des interventions, organisées auparavant dans des séquences. Les séquences de catégories d'interventions résultantes sont ensuite analysées dans un processus d'extraction de motifs séquentiels où des régularités (ou motifs de collaboration) sont extraits. Ces motifs sont présentés à un expert qui les interprète afin de déterminer les mécanismes de collaboration. Au cours de ce chapitre, nous avons également présenté un algorithme approprié pour la réalisation de chacune de ces trois étapes.

Dans le Chapitre 4, nous avons montré comment la méthodologie proposée a été intégrée au sein d'un prototype appelé *historyMiner*. Nous proposons à l'utilisateur un outil complet allant de la visualisation d'un historique des modifications à l'extraction des motifs de collaboration fréquents.

Dans le chapitre 5, nous avons vus deux corpus d'articles de la version anglaise de l'encyclopédie Wikipedia qui ont été particulièrement étudiés. Nous avons présenté et discuté des résultats des expérimentations.

Le nombre restreint des motifs de collaboration fréquents obtenus dans cette étude suscite plusieurs questions comme : les wikistes trouvent-ils inutile de manifester leurs opinions (notamment leur désaccord face à une idée) car ils savent en avance que ça va être révoqué par un autre wikiste ? Plus encore, est-ce sur cette philosophie qu'est basée le succès des sites Wikis et Wikipedia en particulier ? Le consensus sur les Wikis vient-il du fait qu'il n'y a finalement pas de manifestation de désaccord à des idées émises ? Ces questions ont poussé nos collaborateurs du Département de communication à suivre la piste de la page de discussion (associée à un article) et ont pu démontrer que, dans le cas de Wikipedia, cette page de discussion est un outil de coordination que les wikistes utilisent pour résoudre les contradictions avant d'écrire un texte consenti sur la pages d'article et d'avoir par ce fait un article de bonne qualité.

À l'issue de ce travail, nous avons soumis un article intitulé "*Writing a Wikipedia article : Data mining and organizational communication to explain the practices by which contributors maintain the article's coherence*" écrit en collaboration avec Nicolas Bencherki, chercheur au Département de communication de l'Université de Montréal. L'article a été accepté et a fait partie des articles qui ont été présentés à la conférence dédiée à l'ICA (International Communication Association) et qui s'est tenu à Montréal du 22 au 26 mai 2008.

6.2 Perspectives

6.2.1 Utilisation des outils linguistiques perfectionnés

Dans les résultats de cette étude, nous avons mentionné l'utilisation par les wikistes d'un ensemble restreint de connecteurs normalement utilisés dans la langue anglaise pour marquer les relations sémantiques entre les idées. La première question que soulève donc cette étude est : *est-ce que les wikistes utilisent d'autres mots-clés comme des marqueurs nominaux pour exprimer les liens entre leurs interventions et le texte existant ?* Nous pensons qu'un outil linguistique très raffiné, intégrant entre autre un dictionnaire, pourrait être intégré dans l'outil *historyMiner* afin de retrouver (s'il y a lieu) ces autres marqueurs de relations sémantiques et ainsi améliorer la classification sémantique des interventions.
Nous pensons entre autre, aux termes comme "One critic argued" qui est utilisé à la place de "However" et qui n'est pas repérable dans la version actuelle de l'outil.

Un tel outil peut également contribuer à l'amélioration de la taxonomie des interventions. En effet, dans ce travail, nous avons tenu compte de deux aspects des interventions, le syntaxique qui relève des opérations de modification et le sémantique qui s'inspire des relations sémantiques. Un troisième angle des interventions qui serait intéressant à explorer est la reprise d'éléments du texte existant dans les interventions des utilisateurs afin d'assurer la cohérence textuelle. Ce procédé connu sous le nom de récurrence se manifeste, comme nous l'avons vu dans le Chapitre 2, de deux façons. Soit par la reprise directe (la répétition, la pronominalisation, la définitivisation, la substitution lexicale), soit par la reprise référentielle. Bien qu'il soit facile de détecter la reprise directe, la reprise indirecte qui implique l'appartenance à un même champs sémantique nécessite un outil assez sophistiqué.

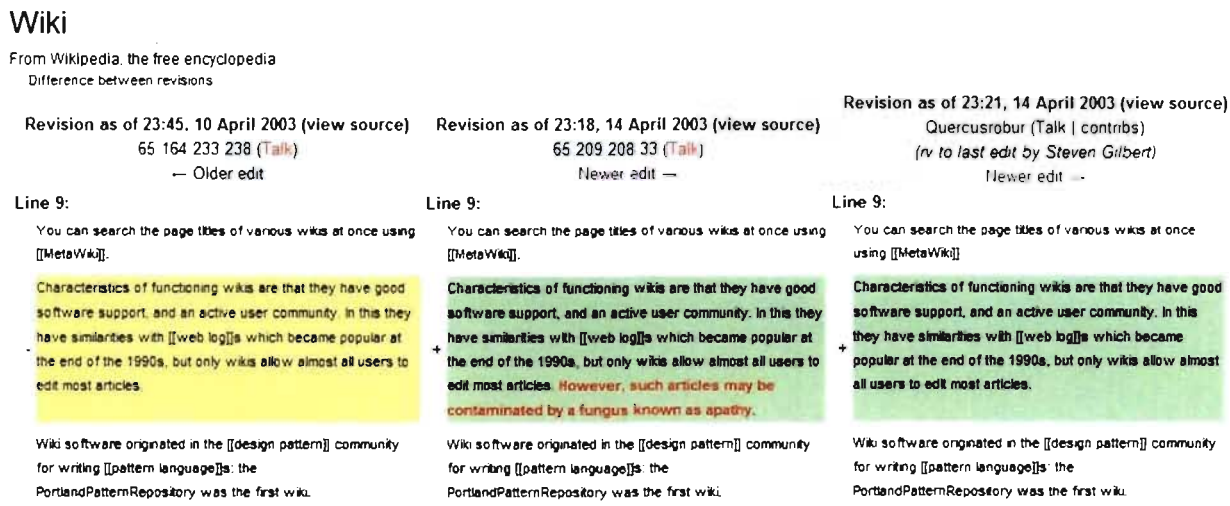

Figure 6.1 – Mécanisme collaboratif de désaccord

6.2.2 Prise en compte du nombre de pages dans la fréquence des motifs de collaboration

Dans l'approche proposée, la fréquence des motifs de collaboration est égal au nombre de paragraphes ayant manifesté ce comportement. Or d'autres motifs de collaboration intéressants sont pénalisés par ce calcul.

Sur la figure 6.1 à la page 83, nous avons la manifestation d'un mécanisme de collaboration que nous pouvons appelé "**Désaccord**" qui se manifeste notamment par l'ajout d'un texte avec un connecteur de restriction et son effacement par après (<(C34)(C46)>). C'est un scénario que nous avons pu identifier dans presque tous les articles dont nous avons analysé mais qui n'apparaît pas dans la liste des motifs de collaboration fréquents car il se manifeste rarement dans l'historique d'une page. Nous pensons que l'introduction d'une notion de fréquence en terme de nombre d'articles ou de pages (au lieu du nombre de paragraphes) ayant manifesté un comportement permettrait de retrouver ce genre de motifs.

BIBLIOGRAPHIE

[1] Rakesh Agrawal et Ramakrishnan Srikant. Mining sequential patterns. Dans Philip S. Yu et Arbee S. P. Chen, éditeurs, *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press. URL `citeseer.ist.psu.edu/agrawal95mining.html`.

[2] Knott Alistair. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Thèse de doctorat, University of Edinburgh, Juillet 1996.

[3] Irena Bellert. On a condition of the coherence of texts. *Semiotica*, pages 335–363, 1970.

[4] Michel Charolles. Introduction aux problèmes de la cohérence des textes (approche théorique et étude des pratiques pédagogiques). *Langue Française*, pages 7–41, 1978.

[5] Norman Fairclough. *Analysing Discourse : Textual Analysis for Social Research*. London : Routledge, 2003.

[6] Le grand dictionnaire terminologique. URL `http://www.granddictionnaire.com/btml/fra/r_motclef/index1024_1.asp`. Consulté le 14 octobre 2007.

[7] M.A.K. Halliday et R. Hasan, éditeurs. *Cohesion in English*. Longman Pub Group, New York, July 1976.

[8] Paul Heckel. A technique for isolating differences between files. *Commun. ACM*, 21(4), 1978.

[9] Micheline Kamber Jiawei Han. *Data Mining : Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[10] Bo Leuf et Ward Cunningham. *The Wiki Way : Quick Collaboration on the Web*. Addison-Wesley, Boston, 2001.

[11] Andrew Lih. Wikipedia as participatory journalism : Reliable sources ? metrics for evaluating collaborative media as a news resource. *5th International Symposium on Online Journalism*, 16-17 April, 2004.

[12] Lita LUNDQUIST, éditeur. *La cohérence textuelle : syntaxe, sémantique, pragmatique.* Nyt Nordisk Forlag : Arnold Busck, Copenhague, 1980.

[13] Rokia Missaoui, Petko Valtchev, Chabane Djeraba et Mehdi Adda. Toward recommendation based on ontology-powered web-usage mining. *IEEE Internet Computing*, 11(4), 2007.

[14] Jim Giles Nature. Internet encyclopaedias go head to head. URL `http://www.nature.com/nature/journal/v438/n7070/full/438900a.html`. Consulté le 1 novembre 2007.

[15] Daniel Terdiman CNET News.com. Study : Wikipedia as accurate as britannica. URL `http://www.news.com/Study-Wikipedia-as-accurate-as-Britannica/2100-1038_3-5997332.html`. Consulté le 1 novembre 2007.

[16] Lorraine Pepin. *La cohérence textuelle.* Groupe Beauchemin éditeur, Laval, 1998.

[17] Ansaf Salleb. *Recherche de motifs fréquents pour l'extraction de règles d'association et de caractérisation.* Thèse de doctorat, Université d'Orléans, Décembre 2003.

[18] Veille scientifique et technologique. Wikibibliographie encyclen. URL `http://wikindx.inrp.fr/biblio_encyclen/`. Consulté le 1 novembre 2007.

[19] Ramakrishnan Srikant et Rakesh Agrawal. Mining sequential patterns : Generalizations and performance improvements. Dans Peter M. G. Apers, Mokrane Bouzeghoub et Georges Gardarin, éditeurs, *Advances in Database Technology - EDBT'96, 5th International Conference on Extending Database Technology,*

Avignon, France, March 25-29, 1996, Proceedings, volume 1057 de *Lecture Notes in Computer Science*. Springer, 1996.

[20] Besiki Stvilia, Michael B. Twidale, Linda C. Smith et Les Gasser. Assessing information quality of a community-based encyclopedia. Dans *IQ*, 2005.

[21] Fernanda B. Viégas, Martin Wattenberg et Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. Dans *CHI '04 : Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004.

[22] Wikimedia META WIKI. Wiki research bibliography. URL `http://meta.wikimedia.org/wiki/Wiki_Research_Bibliography`. Consulté le 1 novembre 2007.

[23] Wikimedia WIKINDX. Wiki research bibliography. URL `http://bibliography.wikimedia.de/index.php`. Consulté le 1 novembre 2007.

[24] Wikipedia. URL `http://www.wikipedia.org/`. Consulté le 14 octobre 2007.

Annexe I

Occurrences des connecteurs dans les articles des deux corpus (1/2)

Connecteur	Podcasting	Civilizations	Connecteur	Podcasting	Civilizations
accordingly	0	11	firstly	0	1
actually	4	130	for	309	4687
all the same	0	0	for a start	0	0
although	11	413	for another thing	0	0
as a consequence	0	4	for example	3	66
as a matter of fact	0	0	for instance	0	14
as a result	0	48	for one thing	0	1
as long as	1	18	furthermore	0	1
assuming that	0	5	given that	0	5
at any rate	0	2	having said that	0	0
at least	3	239	hence	3	89
at once	0	21	however	7	101
at that	5	57	if	42	407
because	32	458	if ever	0	2
beforehand	0	3	if not	2	25
besides	0	28	if only	0	4
but	71	2227	if so	0	1
by the way	0	1	immediatly	0	0
clearly	2	73	in actual fact	0	0
consequently	0	19	in addition	0	38
considering that	0	5	in case	0	8
despite this	0	0	in doing this	0	0
e.g.	23	996	in fact	1	63
even	31	803	in order that	0	1
even so	0	1	in point of fact	0	0
even though	2	43	in so doing	0	0
ever since	0	9	in spite of this	0	0
first	41	1695	in that	9	113
first of all	0	8	in that case	0	1

Annexe II

Occurrences des connecteurs dans les articles des deux corpus (2/2)

Connecteur	Podcasting	Civilizations	Connecteur	Podcasting	Civilizations
in truth	0	1	secondly	0	1
incidentally	0	0	seeing as	0	0
indeed	1	65	so	108	3195
insofar as	0	2	so that	8	120
instantly	0	3	still	20	565
instead	8	140	summing up	0	1
it follows that	0	1	suppose that	0	7
just as	0	36	supposing that	0	2
lastly	0	2	the way	3	71
meanwhile	0	4	then again	0	5
moreover	0	2	thereby	2	25
nevertheless	0	21	therefore	11	127
next	5	136	thirdly	0	0
nonetheless	0	7	this implies that	0	1
now	22	684	this way	2	11
now that	0	16	though	34	1043
obviously	0	16	thus	8	262
on condition that	0	2	to begin with	0	4
on one hand	0	0	to recap	0	0
on the assumption that	0	0	to start with	0	0
on the contrarty	0	0	to sum up	0	0
on the grounds that	0	2	to summurise	0	0
on the one hand	0	1	to the extent that	2	7
on top of this	0	0	to this end	0	2
or rather	2	9	unless	8	33
otherwise	7	53	what is more	0	0
plainly	0	1	whereas	1	48
previously	6	80	yet	2	199
rather	11	293			

Annexe III

Catégorie "Civilizations" - Wikipedia

Category:Civilizations

From Wikipedia, the free encyclopedia

Subcategories

This category has the following 38 subcategories, out of 38 total.

A

- [+] Aegean civilization
- [+] African civilizations
- [+] Akkad
- [+] Amorites
- [+] Ancient Near East
- [+] Andean civilizations
- [+] Aramaeans
- [+] Assyria
- [+] Aztec

B

- [+] Babylonia
- [+] Books about civilizations

C

- [+] China
- [+] Classical civilizations
- [+] Culture by region

E

- [+] Ebla
- [+] Ancient Egypt
- [+] Elam
- [+] Etruscans

F

- [+] Fictional civilizations

G

- [+] Ancient Greece
- [+] Roman Greece

H

- [+] Hattians
- [+] Hittites
- [+] Hurrians

I

- [+] Indus Valley Civilization

L

- [+] Luwians

M

- [+] Maya civilization
- [+] Mesopotamia
- [+] Minoan civilization
- [+] Civilization museums

P

- [+] Phoenicia

R

- [+] Ancient Rome

S

- [+] Sumer

T

- [+] Teotihuacan

U

- [+] Urartu

V

- [+] Vedic civilization
- [+] Ancient Vietnam

W

- [+] Western culture