Processus d'extraction et de validation de la terminologie issue de logs

Hassan Saneifar, Stéphane Bonniol, Anne Laurent, Pascal Poncelet, Mathieu Roche

Processus d'extraction et de validation de la terminologie issue de logs

Hassan Saneifar
Univ. Montpellier 2 - LIRMM
161 rue Ada
Montpellier, France
saneifar@lirmm.fr

Stéphane Bonniol
Satin IP Technologies
Cap Omega, RP B. Franklin,
Montpellier, France
stephane.bonniol@satin-ip.com

Anne Laurent
Univ. Montpellier 2 - LIRMM
161 rue Ada
Montpellier, France
laurent@lirmm.fr

Pascal Poncelet
Univ. Montpellier 2 - LIRMM
161 rue Ada
Montpellier, France
poncelet@lirmm.fr

Mathieu Roche
Univ. Montpellier 2 - LIRMM
161 rue Ada
Montpellier, France
mroche@lirmm.fr

ABSTRACT

Les fichier logs issus des systèmes numériques contiennent des informations importantes concernant les conditions et les configurations de système. Dans le domaine de la conception de circuits intégrés, des fichiers logs sont produits par les outils de conception mais ne sont pas exploités de façon optimale. Les fichiers logs sont des données textuelles multi-source et multi-format qui ont des ont des structures hétérogènes et évolutives. Pour extraire de l'information à partir des logs, la création d'une ontologie du domaine est indispensable. Pourtant, selon les particularités de telles données textuelles, l'application des méthodes classiques de TALN afin d'extraire des termes du domaine qui seront utilisés dans l'ontologie n'est pas une tâche facile. Dans cet article, nous présentons notre approche EXTERLOG qui extrait la terminologie à partir des logs. Ici, nous étudions comment adapter les méthodes du TALN aux logs. Afin d'augmenter la précision des termes extraits, nous les favorisons en leur donnant un score basé sur le Web. Les expérimentations montrent que EXTERLOG obtient des résultats satisfaisants.

Keywords

Traitement du langage naturel, Extraction d'information, Extraction de terminologie, fichiers logs

1. INTRODUCTION

Dans de nombreux domaines d'application, les systèmes numériques produisent automatiquement des rapports. Ces rapports produits, appelés *logs* de système, représentent la source principale d'informations sur la situation des systèmes, des produits ou même des causes des problèmes ayant pu se produire. Dans certains domaines d'application comme les systèmes de conception de Circuits Intégrés, les fichiers logs ne sont pas exploités de façon systématique alors que dans ce domaine particulier, les fichiers logs générés par les outils de conception contiennent les informations essentielles sur les conditions de production et sur les produits finaux.

La création d'une ontologie de domaine est un besoin primordial pour nos futurs travaux sur l'extraction d'informations à partir de fichiers logs. Définir le vocabulaire du domaine est l'une des premières étapes de la construction d'une ontologie. Pour analyser le vocabulaire et la structure lexicale d'un corpus, l'extraction de la terminologie du domaine est l'une des plus importantes phases. Nous avons donc pour but d'extraire la terminologie des fichiers logs. Les termes extraits seront utilisées dans la création d'une ontologie du domaine dans nos travaux futurs. Or, selon les particularités des logs, les méthodes classiques de TALN notamment celles de l'extraction de la terminologie ne sont pas forcement adaptées aux fichiers logs.

Dans ce contexte, un des défis principaux consiste à fournir des approches qui considèrent des structures multi-source, hétérogènes et évolutifs de telles données textuelles ainsi que leurs vocabulaires particuliers. De même, bien que le contenu de ces logs ressemble aux textes écrits en langue naturel (LN), il n'en respecte ni la grammaire ni la structure. Par conséquent, pour extraire la terminologie à partir des logs, nous devons adapter les techniques de traitement automatique du langage naturel (TALN) aux caractéristiques spécifiques de ces données textuelles. Dans cet article, en présentant notre approche EXTERLOG : EXtraction de la TERminologie à partir de LOGs, nous étudions comment adapter les méthodes de TALN aux fichiers logs. EXTERLOG consiste à un pré-traitement et normalisation adaptée aux particularités de telle données textuelles et le processus de l'extraction des co-occurrences (avec et sans utilisation de patrons syntaxiques). Nous avons développé cette phase de l'approche EXTERLOG dans [17]. Toutefois, définir un protocole d'évaluation des termes extraits à partir des logs se rélève indispensable car un autre défi clé dans ce contexte

JFO 2009 December 3-4, 2009, Poitiers, France

est l'évaluation des résultats. Par conséquent, nous définissons une phase de filtrage et validation des termes extrait dans EXTERLOG.

En effet, selon la particularité de ces données, et en raison du bruit élevé dans les résultats, les méthodes d'évaluation classiques ne sont pas forcément pertinentes. Afin de souligner la précision des résultats comme une nécessité du contexte, nous définissons une méthode de filtrage des candidats terminologiques qui sont conformes avec la particularité de ces données. Cette méthode s'appuie sur une mesure de qualité basée sur le Web qui considère à la fois le contexte du domaine.

Ainsi, dans cet article, nous présentons le processus global de l'extraction de terminologie dans les logs notamment le protocole d'évaluation et le filtrage des termes extraits.

Dans la section 2, nous développons l'utilité d'ontologie du domaine et l'extraction de terminologie dans notre contexte ainsi que les particularités et les difficultés de ce contexte. L'approche EXTERLOG est développée dans la section 3. La section 4 présente les différentes expérimentations que nous avons effectuées sur les processus d'extraction et d'évaluation des termes à partir des logs. Enfin, une comparaison de notre méthode EXTERLOG avec le système TERMEXTRACTOR est proposée.

2. CONTEXTE

Aujourd'hui, les systèmes numériques génèrent de nombreux types de fichiers logs, ce qui donne les informations essentielles sur les systèmes. Certains types de fichiers logs, comme les logs issus des systèmes de surveillance de réseau, des interactions des services web ou des logs d'utilisation des sites web (*web usage*) sont largement exploités [21, 10]. Ces types de fichiers logs s'appuient sur la gestion des événements. Cette situation signifie que les informations se trouvant dans ce type de logs sont des événements survenus dans le système qui sont enregistrés chronologiquement. Le contenu de ces logs se conforme à certaines normes grâce à la nature des événements et de leurs utilisations universelles (*e.g.* les services du web).

Au contraire, dans certains domaines comme les systèmes de conception des circuits intégrés, les logs générés sont plutôt des rapports numériques que l'enregistrement de certains événements.

L'objectif de l'exploitation de ces fichiers logs n'est pas d'analyser les événements survenus mais d'extraire de l'information sur l'état et les conditions des produits finaux. Ces fichiers logs sont une source importante d'information pour les systèmes d'information conçus pour interroger et gérer la ligne de production.

2.1 Extraction d'information dans les logs

Afin d'utiliser ces logs dans un système d'information, nous devons implémenter des méthodes d'extraction d'information adaptées aux caractéristiques de ces logs. Les caractéristiques particulières de ces données textuelles rendent peu utilisables les techniques statiques fondées sur l'utilisation simple des schémas d'extraction (e.g. expressions régulières) pour extraire l'information pertinente. Par ailleurs, ces caractéristiques expliquent la raison pour la quelle nous avons besoin d'ontologie du domaine pour extraire d'information à partir des fichiers logs.

Dans la conception des circuits intégrés, il existe plusieurs niveaux. A chaque niveau, plusieurs outils de conception peuvent être utilisés. Cela rend les fichiers logs des données *multi-source*. Malgré le fait que les logs issus du même niveau de conception contiennent les mêmes informations, les structures peuvent significativement différer en fonction de l'outil de conception utilisé. Plus précisément, pour la même information, chaque outil de conception possède souvent son propre vocabulaire. Par exemple, dans l'étape de vérification, nous faisons produire deux fichiers logs (*e.g.* log "A" et log "B") par deux outils de vérification différents. Une information comme le "`Statement coverage`" sera exprimée sous la forme suivante dans le log "A":

TOTAL	COVERED	PERCENT	
Lines	10	11	12
statements	20	21	22

Mais cette même information dans le log "B", sera exprimée par cette simple ligne :

"`EC: 2.1%`"

Tel que montré ci-dessus, la même information, dans deux fichiers logs produits par deux outils différents, est représentée par des structures et un vocabulaire totalement différents. En outre, les outils de conception évoluent au cours du temps et cette évolution se produit souvent de manière imprévisible. Le *format des données* disponibles change, ce qui rend inefficace l'utilisation des méthodes statiques. L'*hétérogénéité* des données traitées existe non seulement entre les fichiers logs produits par différents outils de conception, mais également au sein d'un fichier log donné. Par exemple, les symboles utilisés pour présenter la même notion comme l'entête des tableaux changent au sein d'un log. De même, il existe plusieurs formats pour les ponctuations, les lignes de séparation et la représentation de données manquantes. De plus, certains caractères communs sont utilisés pour présenter différentes notions ou structures. Ainsi, plusieurs lexiques de ce domaine sont constitués de caractères alphanumériques et spéciaux.

Par conséquent, nous avons besoin de méthodes généralisées pouvant être appliquées sur différents logs (*données textuelles multi-format*) qui ont des données multi-format et hétérogènes. Ces méthodes doivent également prendre en compte l'hétérogénéité de la structure et du vocabulaire de ces logs. Pour généraliser au mieux ces schémas d'extraction, nous avons besoin d'ontologie du domaine qui fait la correspondance entre des termes utilisés dans les logs issus des outils différents. Nous utiliserons cette ontologie pour diminuer l'hétérogénéité des termes issus des outils de conception différents. Par exemple, pour vérifier "*l'absence des attributs*" sur les logs, nous devons chercher des phrases différentes dans les logs en fonction de la version et du type d'outil de conception utilisé :

- "`Do not use map_to_module attribute`"
- "`Do not use one_cold or one_hot attributes`"
- "`Do not use enum_encoding attribute`"

Au lieu d'utiliser plusieurs patrons chacun adapté à une phrase, en associant les termes "`map_to_module attribute`", "`one_hot attributes`" et "`enum_encoding attribute`" au concept "*Absence d'attributs*", nous utiliserons un patron général qui s'adapte automatiquement aux différents logs en utilisant l'ontologie du domaine. Plusieurs approches s'appuient également sur les ontologies de domaine pour mieux guider

des démarches d'extraction d'information [9]. Dans notre contexte, nous utiliserons l'ontologie du domaine pour diminuer l'hétérogénéité des termes issus des outils de conception différents. Cette ontologie présente les termes correspondants dans les différents logs issus des outils différents. Cela nous permet également de catégoriser des termes associés à un concept cherché sur les logs.

La création d'ontologie nécessite tout d'abord une analyse lexicale de corpus afin identifier les termes du domaine [18]. Nous cherchons donc à identifier la terminologie propre des logs de chaque outil de conception. Nous étudierons ensuite les termes de chaque outil pour faire la correspondance entre aux et créer l'ontologie du domaine. Ainsi, dans cet article, nous avons pour objectif d'étudier l'extraction de la terminologie à partir de fichiers logs. Nous insistons ici sur le fait que dans notre contexte, l'ontologie du domaine doit être créé directement et en s'appuyant principalement sur le corpus de logs car l'extraction d'information sera effectuée directement sur le corpus de logs. Cela veut dire qu'une terminologie extraite à partir d'autres documents du domaine comme les références de conception n'est pas complète ou pertinente car plusieurs termes et structures lexicales sont utilisés uniquement dans les logs. Notons que nous cherchons à créer plutôt un réseau sémantique de concepts car nous ne cherchons pas à ce stade des hiérarchies de concepts.

Par ailleurs, le langage utilisé dans ces logs est une difficulté qui influence aussi des méthodes d'extraction d'information. Bien que la langue utilisée dans ces logs soit l'anglais, les contenus de ces logs n'en respectent pas la grammaire "classique". Dans cet article, nous étudions donc ces méthodes et leur pertinence dans ce contexte spécifique. Finalement, nous proposons une approche d'extraction de terminologie EXTERLOG adaptée à ces particularités.

2.2 Méthodes d'extraction de la terminologie

L'extraction de la terminologie de domaine dans les données textuelles est une tâche essentielle afin d'établir des dictionnaires spécialisés du domaine [14]. L'extraction des bigrammes[1] est une étape importante pour identifier des candidats terminologiques. Nous cherchons des bigrammes dans les fichiers logs avec deux approches différentes :

1. extraction des bigrammes en utilisant des patrons syntaxiques,
2. extraction des bigrammes sans utilisation de patrons syntaxiques.

L'extraction de la terminologie de domaine dans les données textuelles est une tâche essentielle afin d'établir des dictionnaires spécialisés du domaine [14]. Les bigrammes sont utilisés dans [13] comme des attributs (*features*) pour améliorer la performance de la classification de textes. Les séries de trois mots (*i.e.* trigrammes) ou plus ne sont pas toujours essentiels [11].
Les règles et les grammaires définies sont utilisées dans [7] afin d'extraire les termes nominaux ainsi que pour les évaluer. Les méthodes d'apprentissage automatique fondées sur les modèles cachés de Markov (HMMs) sont utilisées dans [4] pour extraire la terminologie dans le domaine de biologie moléculaire. EXIT, présenté par [14], est une approche itérative qui trouve les termes de façon incrémentale. Un terme découvert à une itération est utilisé dans la prochaine itération afin de trouver des termes plus complexes. Certains travaux tentent d'extraire les co-occurrences dans une fenêtre de taille fixe (*normalement cinq mots*). Dans ce cas, les mots extraits peuvent ne pas être directement liés [12]. XTRACT évite ce problème en considérant les positions relatives des co-occurrences. XTRACT est un système d'extraction de terminologie, qui identifie des relations lexicales dans les corpus volumineux de textes anglais [19]. SYNTEX, présenté par [1], effectue l'analyse syntaxique des textes pour identifier les noms, les verbes, les adjectifs, les adverbes, les syntagmes nominaux et les syntagmes verbaux. Il analyse les textes en appliquant des règles syntaxiques pour en extraire les termes. TERMEXTRACTOR, présenté par [18], est un logiciel pour l'extraction des termes pertinents dans un domaine spécifique. L'application prend en entrée un corpus de documents de domaine, analyse les documents, et extrait des termes syntaxiquement plausibles. Afin de sélectionner uniquement les termes qui sont pertinents pour le domaine, certaines mesures fondées sur l'entropie sont utilisées.

Les méthodes statistiques sont généralement utilisées associées à des méthodes syntaxiques pour évaluer la pertinence des candidats terminologiques [6]. Ces méthodes sont fondées sur des mesures statistiques pour valider comme terme pertinent un candidat extrait. Parmi ces mesures, la fréquence d'occurrences des candidats est une notion de base. Or, ces méthodes statistiques ne sont pas pertinentes dans notre contexte. En effet, les approches statistiques peuvent faire face aux termes ayant une fréquence élevée dans le corpus, mais ont tendance à manquer des termes peu fréquents [8]. Selon les fichiers logs décrits ci-dessus, la répétition des mots est plus rare. Chaque partie d'un log contient certaines informations indépendantes d'autres parties.

Evaluation des termes basées sur d'autres ressources telles que le Web est étudiée par de nombreux travaux. Le Web, comme un vaste corpus, est de plus en plus utilisés dans les méthodes de TALN spécialement pour la validation des résultats. Toutefois, dans notre contexte, nous étudions le corpus d'un domaine très *spécialisé*. Les termes utilisés dans ce domaine sont les termes spécialisés et pas souvent vus sur le Web. Ensuite, comme montré ci-dessus, nous ne pouvions pas utiliser les mesures statistiques classiques basées sur la fréquence des termes dans le corpus afin de donner un score à chaque terme extrait. Par conséquent, notre approche vise à réduire le bruit dans les résultats, soulignant ainsi la précision, en filtrant les termes extraits en fonction d'une mesure basé sur le Web qui considère en même temps le contexte spécialisé des logs.

Beaucoup de travaux comparent les différentes techniques d'extraction de terminologie et leur performance. Mais la plupart de ces travaux sont expérimentés sur les données textuelles, qui sont les textes classiques écrits en langage naturel. La plupart des corpus utilisés sont structurés de manière cohérente. En particulier, ces données textuelles suivent la grammaire des langues naturelle. C'est la raison pour la quelle les méthodes classiques d'extraction de termes ne sont pas pertinentes dans notre contexte par rapport aux caractéristiques particulières des logs (cf. section 2). Cela

[1] un n-gramme de mots est défini comme une série de "n" mots.

impose une adaptation particulière pour que ces méthodes soient pertinentes dans le cas des logs. Par conséquent, une comparaison expérimentale des différents outils à partir des fichiers logs ne se révèle pas intéressante.

Dans la prochaine section, nous expliquons l'approche que nous proposons pour extraire la terminologie dans les fichiers logs.

3. EXTERLOG : EXTRACTION DE LA TERMINOLOGIE À PARTIR DE LOGS

Nous allons expliquer notre approche EXTERLOG développée afin d'extraire la terminologie dans les fichiers logs. Le processus d'extraction consiste à la normalisation, le prétraitement des fichiers logs et leur étiquetage grammatical afin d'en extraire les termes. EXTERLOG contient également une phase de filtrage des termes extraits en fonction d'une mesure de scoring basé sur le Web.

3.1 Pré-traitement & normalisation

L'hétérogénéité des fichiers logs est une difficulté considérable qui peut influencer la performance des méthodes d'extraction d'information. Afin d'avoir des patrons d'extraction d'information généralisés, nous effectuons une série de prétraitements et de normalisation sur les logs.
Compte tenu des spécificités de nos données textuelles, nous appliquons une méthode de normalisation adaptée aux logs pour rendre le format et la structure de logs plus cohérents. Nous remplaçons les ponctuations, les lignes de séparation et les en-têtes des tableaux par des caractères spéciaux pour réduire l'ambiguïté. Puis, nous segmentons les logs, considérant que certains mots ou structures ne doivent pas être segmentés. Par exemple, le mot technique "`Circuit4-LED3`" est un mot unique qui ne doit pas être segmenté en deux mots "Circuit4" et "LED3". De plus, nous identifions les lignes représentant l'en-tête des tableaux pour distinguer les lignes de séparation. Après normalisation des logs, nous avons moins d'ambiguïté ou de symboles communs pour différents concepts. Cette normalisation rend la structure des logs issus des différents outils plus homogène.

3.2 Étiquetage grammatical

L'étiquetage grammatical est une méthode classique du TALN pour analyser les fichiers de textes de façon à réaliser l'annotation grammaticale des mots. Dans notre contexte, selon la nature des fichiers logs, il existe des difficultés et des limites pour appliquer un étiqueteur grammatical sur de telles données textuelles.

En effet, les techniques classiques sont développées selon la grammaire standard de la langue naturelle. En outre, elles sont normalement entraînées sur des textes écrits dans un langage naturel standard comme les journaux. Par exemple, une phrase se finit par un point, ce qui n'est pas le cas dans les fichiers logs que nous traitons. Plus précisément, dans ces derniers, les phrases et les paragraphes ne sont pas toujours bien structurés. De plus, il existe plusieurs constructions qui ne vérifient pas les structures de phrases de la langue naturelle.

Pour identifier le rôle grammatical des mots dans les logs, nous employons l'étiqueteur grammatical BRILL [2]. Nous avons adapté BRILL au contexte des logs en utilisant de nouvelles règles *contextuelles* et *lexicales*. Par exemple, un mot commençant par un nombre est considéré comme un "cardinal" par l'étiqueteur de BRILL. Or, dans les fichiers logs, il existe de nombreux mots comme `12.1vSo10`; qui ne devraient pas être étiquetés comme "cardinal". C'est pourquoi nous avons défini des règles lexicales et contextuelles spécifiques à nos données.

Les structures particulières des logs peuvent être une information importante pour extraire les schémas dans nos travaux futurs. Par conséquent, nous maintenons la structure des textes inchangée lors de l'étiquetage des logs. Pour cela, nous présentons les nouvelles étiquettes comme "\TH", qui représente la ligne utilisée dans l'en-tête des tableaux ou "\SPL" pour les lignes de séparation. Ces étiquettes, que nous appelons "*les étiquettes de structure de document*", peuvent être identifiées lors de la normalisation et par des règles contextuelles/lexicales spécifiques qui ont été introduites dans BRILL. Nous obtenons finalement les logs étiquetés par les rôles grammaticaux des mots et également par les étiquettes qui déterminent la structure des logs.

3.3 Extraction des bigrammes

Afin d'identifier les co-occurrences dans les logs, nous considérons deux solutions :

1. extraction des bigrammes en utilisant des patrons syntaxiques (ci-après "*Bigrammes-AP*"[2]),

2. extraction des bigrammes sans utilisation de patrons syntaxiques (ci-après "*Bigrammes-SP*"[3]).

Dans la première, nous utilisons le filtrage de mots par des patrons syntaxiques. Les patrons syntaxiques déterminent les mots adjacents ayant les rôles grammaticaux définis. Les patrons syntaxiques que nous utilisons pour extraire les bigrammes-AP sont :
- "\AJ - \NN" (Adjectif-Nom)
- "\NN - \NN" (Nom-Nom)

Notons que selon nos expérimentations et le contenu des logs, les patrons syntaxiques complexes ou les termes composés de plus de deux mots n'ont pas l'utilité dans le contexte de fichiers logs. Ainsi, les algorithmes d'extraction de terminologie plus complexes, comme les algorithmes qui extraient les terminologies dans une fenêtre de mots ne sont pas pertinents dans ce contexte car il existe peu de contenus textuelles écrits en langue naturelle et ils sont systématiquement séparés par des tableaux ou les chiffres.

Au contraire, dans la deuxième solution, l'extraction des bigrammes-SP (sans utilisation des patrons syntaxiques) ne dépend pas du rôle grammatical des mots. Afin d'extraire les bigrammes-SP significatifs, nous considérons le bruit existant dans les logs. Par conséquent, nous normalisons et segmentons les logs pour diminuer le taux de bruit. Les bigrammes extraits représentent deux mots adjacents "ordinaires".

Ces termes extraits à ce stade doivent être évalués et filtrés afin de favoriser les termes les plus pertinents du domaine.

3.4 Filtrage des termes

Tous les termes extraits ne sont pas nécessairement les termes pertinents du domaine. En raison de certains fichiers logs volumineux et le grand vocabulaire des logs, il existe

[2]AP : Avec Patron
[3]SP : Sans Patron

plusieurs termes extraits. En outre, à cause des caractéristiques particulières de ces données, en dépit des méthodes adaptées de la normalisation et l'étiquetage que nous avons utilisés, il existe des termes non-pertinents parmi les termes extraits. En outre, nous sommes axés sur un domaine spécialisé où seulement certains termes sont vraiment des termes propres du domaine. Ainsi, nous évaluons et filtrons les termes extraits dans le but de favoriser les termes les plus pertinents selon le contexte. Les mesures statistiques sont souvent utilisées dans l'extraction de terminologie pour évaluer les termes. Les mesures suivants sont les plus largement utilisés.

Information Mutuelle. Une des mesures les plus utilisées pour calculer une sorte de relation entre les co-occurrences est l'information mutuelle (IM) [3]. Cette mesure se calcule selon la formule simplifiée suivante, où nb désigne le nombre d'occurrences :

$$IM(x,y) \quad = \quad \log_2 \frac{nb(x,y)}{nb(x)nb(y)}$$

Information Mutuelle Cube. L'information mutuelle cube est une mesure empirique fondée sur IM, qui renforce l'impact de la fréquence de co-occurrences, ce qui est absent dans l'original IM [5].

$$IM3(x,y) \quad = \quad \log_2 \frac{nb(x,y)^3}{nb(x)nb(y)}$$

Cette mesure est utilisée dans plusieurs travaux liés l'extraction des termes dans les textes [15].

Coefficient de Dice. Le coefficient de Dice est une mesure de qualité intéressante [20]. Elle se calcule en fonction de la fréquence d'occurrence :

$$Dice(x,y) \quad = \quad \frac{2 \times nb(x,y)}{nb(x) + nb(y)}$$

Ces mesures sont fondées sur la fréquence d'occurrences des termes dans corpus. Or, le scoring des termes basé sur les fréquences des termes dans les logs n'est pas une approche pertinente dans notre contexte. Comme nous l'avons déjà expliqué, les techniques basées sur les fréquences des termes dans un corpus ne sont pas pertinents dans ce contexte, car les termes *représentatifs* n'ont pas forcement une haute fréquence dans les fichiers logs. C'est pourquoi nous effectuons le scoring des termes en fonction de leur fréquence sur le Web comme un vaste corpus où la fréquence d'un terme peut être représentatif. Toutefois, travaillant sur un domaine spécialisé, nous obtenons des scores biaisés en simplement comptant les fréquences d'occurrences des termes sur le Web. En effet, sur le Web, on capte les occurrences de termes, quel que soit le *contexte* dans lequel elles sont vues. Ainsi, il convient de ne considérer que les occurrences des termes sur le Web qui sont situés dans le contexte de la conception de CI. Nous utilisons donc une extension des mesures décrites ci-dessous appelée $AcroDef$. $AcroDef$ est une mesure où le contexte et les ressources du Web sont des caractéristiques essentielles à prendre en compte (voir [15]). Les formules ci-dessous se définissent les différents types de mesures $AcroDef$, respectivement basées sur IM et IM3.

$$AcroDef_{IM}(a^j) = \frac{nb(\bigcap_{i=1}^{n} a_i^j + C)}{\prod_{i=1}^{n} nb(a_i^j + C | a_i^j \notin M_{stop-words})} \quad \text{où } n \geq 2$$

$$AcroDef_{IM3}(a^j) = \frac{nb(\bigcap_{i=1}^{n} a_i^j + C)^3}{\prod_{i=1}^{n} nb(a_i^j + C | a_i^j \notin M_{stop-words})} \quad \text{où } n \geq 2$$

Dans $AcroDef$, le *contexte* "C" est représenté comme un ensemble des mots significatifs. La fonction nb utilisée dans les mesures précédentes représente le nombre des pages retournées par le moteur de recherche en fonction d'un terme donné comme la requête. Ainsi, $nb(a_i^j + C)$ retourne le nombre de page en utilisant $a_i^j + C$ comme la requête ce qui signifie tous les mots de terme a^j en plus des mots representant le contexte C.

Par exemple, pour un terme comme "atpg patterns" constitué de deux mots ($i = 2$), $nb(atpg \bigcap patterns + C)$ est le nombre des pages retournées en envoyant la requête "atpg pattern" AND C sur un moteur de recherche où C est l'ensemble des mots représentant le contexte. La formule d'$AcroDef_{Dice}$ basé sur la formule de Dice est représentée ainsi :

$$\frac{\left|\{a_i^j + C | a_i^j \notin M_{stop-words}\}_{i \in [1,n]}\right| \times nb(\bigcap_{i=1}^{n} a_i^j + C)}{\sum_{i=1}^{n} nb(a_i^j + C | a_i^j \notin M_{stop-words})} \quad \text{où } n \geq 2$$

Dans [15], "C" est introduit comme un ensemble des mots significatifs (*e.g.* chiffrage, information et code pour représenter le contexte Cryptographie). Les auteurs utilisent l'opérateur "AND" des moteurs de recherche entre les mots définissant le contexte. Ceci revient à retourner les pages contenant tous les mots de "C".

Pour préciser les mots qui représentent le contexte des logs, nous construisons un corpus de documents comprenant les documents de référence des outils de conception de circuits intégrés ainsi que trois documents de domaine proche. Nous classons les mots du corpus à l'aide de la mesure *tf-idf* (voir [16]). Tf-idf donne plus de poids aux mots fréquents d'un domaine qui ne sont pas fréquents dans les autres. Ensuite, nous avons choisi les cinq premiers mots (*classés dans l'ordre donné par le tf-idf*) issus des documents de conception de CI comme les mots représentant le contexte.

Tel que souligné ci-dessus, nous recherchons des pages Web contenant un terme donné et en même temps les mots du contexte (*en utilisant les opérateurs AND et OR entre les mots du contexte*). Enfin, les termes extraits sont classés selon leurs scores $AcroDef$. Nous favorisons les termes les mieux classés en filtrant les termes ayant les scores $AcroDef$ les plus aibles.

4. EXPÉRIMENTATIONS

Les expérimentations consistent en trois phases. Nous expérimentons d'abord les deux méthodes utilisées pour extraire la terminologie à partir des logs : (1) en utilisant

des patrons syntaxiques (*bigrammes-AP*) et (2) sans utilisation des patrons syntaxiques (*bigrammes-SP*). Ensuite, nous validons notre protocole d'évaluation automatique des termes extraits. Enfin, une comparaison de notre méthode EXTERLOG avec le système TERMEXTRACTOR est proposée.

Dans toutes les expérimentations, le corpus de logs d'une taille de 950 Ko est constitué des logs de tous les niveaux de conception.

4.1 Evaluation automatique des bigrammes

Pour analyser la performance des deux approches choisies pour l'extraction des bigrammes, nous devons évaluer les termes extraits. Pour évaluer de manière automatique leur pertinence, nous comparons les bigrammes-AP et bigrammes-SP aux termes extraits à partir des documents de référence du domaine. Pour chaque niveau de conception des circuits intégrés, nous utilisons certains documents, qui expliquent les principes de la conception et particulièrement les détails des outils de conception. Nous employons ces documents comme "références expertes" dans le cadre d'une validation automatique. En effet, si un terme extrait des logs est utilisé dans les références du domaine, nous pouvons le considérer comme un terme valide du domaine. Pourtant, il existe plusieurs termes propres aux logs surtout les termes techniques qui ne sont pas utilisés dans le corpus de référence. C'est pourquoi une validation par un expert, effectuée dans la section 4.2, est indispensable pour compléter la validation automatique.

Le corpus de documents de référence est composé d'environ trois documents par niveau de conception. Ces documents sont de taille considérable. Chaque document est constitué d'environ 600 pages. Étant donné que le corpus de référence est constitué des textes écrits en langue standard contrairement aux logs, nous appliquons la méthode classique d'extraction de terminologie pour extraire les termes à partir du corpus de référence.

Niveau 1		Niveau 2		Niveau 3		Niveau 4		Niveau 5	
AP	SP	AP	SP	AP	SP	AP	SP	AP	SP
67,7	11,3	20,7	6,5	37,8	9,9	40,1	6,5	19,6	5,1

Table 1: Proportion de bigrammes-AP et de bigrammes-SP retrouvés dans les documents de références

Nous calculons la proportion P des bigrammes-AP et SP retrouvés dans les documents de références.

$$P = \frac{\mid Bigrammes \cap Termes\,de\,références \mid}{\mid Bigrammes \mid}$$

Le tableau 1 montre la proportion des bigrammes-AP et SP. En effet, cette proportion donne une tendance générale quant à la qualité des termes extraits par notre système. Notons que pour évaluer la pertinence des termes extraits, il faudrait demander à un expert d'évaluer manuellement tous les termes proposés par EXTERLOG. C'est la raison pour laquelle nous avons défini le protocole d'évaluation des termes extraits en utilisant $AcroDef$. Nous allons expérimenter ce protocole dans la section suivante.

La comparaison des bigrammes-AP et bigrammes-SP relativement aux termes de références montre que l'extraction de la terminologie fondée sur les patrons syntaxiques est tout à fait pertinente sur les données logs. Malgré le fait que la normalisation et l'étiquetage des données logs ne soient pas une tâche facile, nos expérimentations montrent qu'un effort dans ce sens est tout à fait utile dans le but d'extraire une terminologie de qualité.

4.2 Evaluation manuelle des bigrammes

La validation des termes par un expert est une tâche difficile en raison du nombre des termes extraits par EXTERLOG. Nous effectuons donc les expérimentations sur un échantillon des termes. Cet échantillon est composé des 700 termes (bigrammes-AP) les plus fréquents dans le corpus de logs.

Les mots représentant le contexte dans $AcroDef$ sont déterminés de la manière expliquée dans la section 3.4. Nous utilisons le moteur de recherche "Google" pour récupérer le nombre de pages web contenant à la fois un terme donné et *deux* ou *plus* mots du contexte. Supposons "CPU time" comme un terme donné où C_i $i \in \{1-5\}$ sont les 5 mots représentant le contexte. La requête envoyée à Google est "CPU time" AND C_1 AND C_2 OR C_3 OR C_4 OR C_5.

Une fois $AcroDef$ calculé, nous classons les termes en fonction de leurs scores $AcroDef$. Plus la valeur d'$AcroDef$ est *élevée*, plus le terme est *représentatif* dans notre contexte. Ainsi, nous favorisons les termes les mieux classés dans le but d'augmenter la précision en diminuant le taux de bruit (les termes non pertinents) dans les résultats.

Pour évaluer la performance d'$AcroDef$ en tant que mesure de qualité, nous faisons valider les termes par deux experts du domaine. Un des experts a annoté les termes comme "pertinent" ou "non pertinent" et l'autre a confirmé les annotations du premier expert. Ensuite, nous évaluons la fonction de rang utilisée pour classer les termes (*i.e.* $AcroDef$) en utilisant les courbes ROC (Receiver Operating Curve).

Une courbe ROC permet de comparer des algorithmes qui classifient des éléments d'un jeu de données en deux groupes *positif* et *négatif*. Elle indique la capacité de la fonction de rang (ici $AcroDef$) à placer les positifs devant les négatifs. Dans notre cas, cela consiste à placer les termes pertinents devant les termes non pertinents en les ordonnant en fonction des scores d'$AcroDef$. Un score efficace doit conduire à des distributions bien séparées. En utilisant des courbes ROC, nous évaluons à quel point $AcroDef$ est pertinent comme une mesure pour distinguer les termes positifs et négatifs. Sur une courbe ROC, plus elle est proche de la partie supérieure du carré (sur le diagramme), meilleure est la séparation. Lorsque les deux densités sont identiques, la courbe ROC se confond avec la diagonale du carré.

Il existe également un indicateur synthétique dérivé à partir de la courbe ROC. Il s'agit de l'AUC (Area Under Curve) qui est la surface entre la courbe et l'axe des abscisses. Cela indique la probabilité d'un individu (*dans notre cas un terme*) positif d'être classé devant un individu négatif. Il existe une valeur seuil, si l'on classe les individus au hasard, l'AUC sera égale à 0.5. La figure 1 montre les courbes ROC obtenues en fonction d'$AcroDef_{IM}$, $AcroDef_{IM3}$ et $AcroDef_{Dice}$.

Le tableau 2 présente l'AUC calculée en considérant les m meilleurs termes classés avec nos trois fonctions de rangs

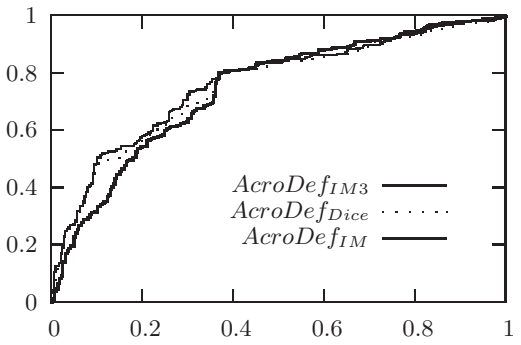

Figure 1: courbes ROC en fonction des trois types d'*AcroDef*

m	AUC_{MI}	AUC_{MI3}	AUC_{Dice}
200	0.53	0.60	0.59
300	0.61	0.70	0.70
400	0.62	0.67	0.64
500	0.66	0.72	0.71
600	0.72	0.75	0.75
700	0.74	0.77	0.76

Table 2: AUC obtenue à chaque niveau de filtrage en fonction des courbes ROC de la fonction de rang *AcroDef*

$AcroDef_{IM}$, $AcroDef_{IM3}$, $AcroDef_{Dice}$. Ce classement est effectué à partir de l'échantillon des 700 termes les plus fréquents extraits dans le corpus de logs.

Les résultats montrent que la mesure *AcroDef* basée sur IM3 classe mieux les termes extraits en fonction de leur pertinence.

5. CONCLUSION & PERSPECTIVES

Dans cet article, nous avons décrit un type particulier de données textuelles : les fichiers logs générés par des outils de conception de circuits intégrés. Étant donné que ces fichiers logs sont des données textuelles multi-source, multi-format, hétérogènes et évolutives, les méthodes de TALN et extraction d'information ne sont pas nécessairement adaptées afin d'extraire de l'information. Pour extraire la terminologie des logs, nous avons extrait des co-occurrences. Pour cela, nous avons adapté les méthodes de prétraitement, de la normalisation et de l'étiquetage grammatical aux particularités des logs. Les résultats montrent que l'étiquetage grammaticale des logs, malgré le fait qu'ils ne sont pas des données textuelles classiques, est une approche pertinente afin d'extraire de l'information. Dans l'objectif de favoriser les termes les plus pertinent du domaine, nous avons défini un protocole d'évaluation automatique des termes extraits qui utilise une mesure basée sur le Web en considérant en même temps le contexte spécialisé des logs. Les expérimentations montrent que notre approche de l'extraction de la terminologie à partir des logs, EXTERLOG, est une approche adaptée à de telles données textuelles.

6. REFERENCES

[1] D. Bourigault and C. Fabre. Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire - Université Toulouse le Mirail*, (25):131–151, 2000.

[2] E. Brill. A simple rule-based part of speech tagger. In *In Proceedings of the Third Conference on Applied Natural Language Processing*, pages 152–155, 1992.

[3] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Computational Linguistics*, volume 16, pages 22–29, 1990.

[4] N. Collier, C. Nobata, and J. Tsujii. Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. *Journal of Terminology, John Benjamins*, 7(2):239–257, 2002.

[5] B. Daille. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Universit Paris 7, 1994.

[6] B. Daille. Conceptual structuring through term variations. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 9–16, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[7] S. David and P. Plante. De la nécessité d'une approche morpho-syntaxique en analyse de textes. *Intelligence Artificielle et Sciences Cognitives au Québec*, 2(3):140–155, September 1990.

[8] D. A. Evans and C. Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 17–24, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[9] F. Even and C. Enguehard. Extraction d'informations à partir de corpus dégradés. In *Proceedings of 9ème conference sur le Traitement Automatique des Langues Naturelles (TALN'02)*, pages 105–115, 2002.

[10] F. M. Facca and P. L. Lanzi. Mining interesting knowledge from weblogs: a survey. *Data Knowl. Eng.*, 53(3):225–241, 2005.

[11] M. Grobelnik. Word sequences as features in text-learning. In *In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98*, pages 145–148, 1998.

[12] D. Lin. Extracting collocations from text corpora. In *In First Workshop on Computational Terminology*, pages 57–63, 1998.

[13] C. meng Tan, Y. fang Wang, and C. do Lee. The use of bigrams to enhance text categorization. In *Inf. Process. Manage*, pages 529–546, 2002.

[14] M. Roche, T. Heitz, O. Matte-Tailliez, and Y. Kodratoff. EXIT: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In *Proceedings of JADT'04 (International Conference on Statistical Analysis of Textual Data)*, volume 2, pages 946–956, 2004.

[15] M. Roche and V. Prince. AcroDef : A quality measure for discriminating expansions of ambiguous acronyms. In *CONTEXT*, pages 411–424, 2007.

[16] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA, 1987.

[17] H. Saneifar, S. Bonniol, A. Laurent, P. Poncelet, and M. Roche. Terminology extraction from log files. In *DEXA '09: Proceedings of the 20th international conference on Database and Expert Systems Applications*. Springer-Verlag, 2009.

[18] F. Sclano and P. Velardi. Termextractor: a web application to learn the shared terminology of emergent web communities. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*, Funchal, Portugal, 2007.

[19] F. Smadja. Retrieving collocations from text: Xtract. *Comput. Linguist.*, 19(1):143–177, 1993.

[20] F. Smadja, K. R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

[21] K. Yamanishi and Y. Maruyama. Dynamic syslog mining for network failure monitoring. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 499–508, New York, NY, USA, 2005. ACM.