Vers des méta-règles de contexte appréciées par la IIE pour la RI

Sourour Belhaj Rhouma * Chiraz Latiri ** Yahya Slimani *

* Laboratoire LISI, INSAT, Université de Carthage, Tunisia
sourour.bhr@gmail.com, yahya.slimani@gmail.com
** Laboratoire LIPAH, Faculté des Sciences de Tunis,
Campus Universitaire Tunis El Manar, 1060 Tunis, Tunisie
chiraz.latiri@gnet.tn

RÉSUMÉ. *Le processus de Fouille de Textes (FT), basé sur l'extraction des règles d'association en utilisant un algorithme, génère une quantité importante de règles d'association. Dans cet article, ce sont des règles d'association non redondantes résultantes d'un processus d'extraction à partir d'un corpus de textes. Nous proposons tout d'abord de montrer l'intérêt et l'utilité de règles d'association filtrées par une mesure de qualité autre que la confiance, en particulier l'Intensité d'Implication Entropique (IIE). D'autre part, nous présentons une nouvelle structure de données "les Méta-règles de contexte" (MR) afin de réduire la base de règles en mettant en valeur les nouvelles mesures correspondantes. Nous étudions nos approches dans le cadre de la recherche d'information (RI) pour l'expansion de requêtes (*QE*). Nos expériences ont été menées sur une collection en anglais de* CLEF 2003. *Les résultats ont montré une amélioration significative de la pertinence système.*

ABSTRACT. *Text Mining (TM) process, based on mining association rules using an algorithm, generates a significant amount of association rules. In this article, these are non-redundant association rules resulting from a mining process from a text corpus. We first propose to show the interest and usefulness of association rules filtred by a quality measure other than confidence, especially Entropic Implication Intensity (EII). Secondly, we present a new data structure "Contextual meta-rules" to reduce the rule base by highlighting the new corresponding measures. We study our approachs in the context of Information Retrieval (IR) through the automatic Query Expansion (QE). Our experiments was conducted on English collection of* CLEF 2003. *The results showed a significant improvement.*

MOTS-CLÉS : *FT, règles d'association entre termes, IIE, Méta-règles de contexte, QE, RI*

KEYWORDS: *TM, association rules between terms, EII, Contextual meta-rules, QE, IR*

CORIA 2015, pages 1 à 16

1. Introduction et motivations

La RI étudie le processus d'adéquation entre la requête d'un utilisateur et une collection de documents. Le résultat de ce processus est souvent un sous-ensemble de documents pertinents contenant les mêmes termes de la requête originelle. Le modèle classique de RI (Salton et McGill, 1983) consiste à attribuer, à chaque document d'une collection, des termes d'indexation, dits *index* du document. Ce modèle limite les requêtes à l'ensemble global des termes de l'index, et utilise des mesures de correspondance entre les requêtes et les documents.

Une des difficultés rencontrées au cours d'une session de recherche documentaire est liée aux choix des termes d'interrogation. Afin d'avoir des documents pertinents, l'utilisateur est contraint d'utiliser le "*vocabulaire de description du document*" propre au système. Face à cette contrainte, il est possible de faire appel à la technique *d'expansion de requêtes* (Buckley *et al.*, 1994) afin d'améliorer la correspondance requête/document. Cette technique consiste à étendre la requête par des termes additionnels, corrélés à ceux de la requête originelle. Intuitivement, l'apport d'une telle technique ne se réduit pas à l'amélioration du rappel en récupérant des documents pertinents qui ne peuvent pas être trouvés par la requête utilisateur, mais également à améliorer la précision des documents trouvés en les plaçant en haut de la liste des documents pertinents.

L'issue de recherche que nous suggérons est le déploiement des règles d'association entre termes (Agrawal et Skirant, 1994) dans un processus d'expansion de requêtes en RI (Haddad *et al.*, 2000 ; Lin *et al.*, 2008 ; Rungsawang *et al.*, 1999 ; Tangpong et Rungsawang, 2000). Toutefois, face au nombre très important de règles d'association entre termes qui peuvent être découvertes à partir d'une collection de documents, dans (Latiri *et al.*, 2012), les auteurs ont proposé un nouveau processus d'expansion automatique de requêtes moyennant la base générique minimale de règles d'association non redondantes et qui sont appréciées uniquement par la mesure de *confiance* (Latiri *et al.*, 2012). Les évaluations expérimentales sur un ensemble de collections de test ont montré une amélioration significative de la pertinence de la tâche RI.

Néanmoins, l'évaluation de la qualité des règles d'association entre termes utilisées en RI s'est limitée dans les travaux de l'état de l'art à la mesure de la confiance. Nous nous intéressons dans ce travail de recherche à explorer l'impact de l'utilisation d'une mesure de qualité telle que *l'Intensité d'Implication Entropique (IIE)*. Notre motivation est double : en premier lieu, il s'agit d'étudier l'utilité des règles d'association entre termes appréciées par la IIE dans l'expansion de requêtes. En deuxième lieu, nous proposons une nouvelle structure de données dans le but de réduire la base des règles et les déployer dans le processus d'expansion de requêtes.

La suite de l'article est organisé comme suit : la deuxième section présente une revue de la littérature des travaux reliés à l'expansion de requêtes. La section 3 est dédiée à la définition et la génération des règles d'association entres termes selon la IIE. Nous présentons dans la section 4 "les Méta-règles de contexte" pour réduire la

base des règles. Dans la section 5, nous détaillons le processus du déploiement des règles d'association entre termes et des méta-règles de contexte pour l'expansion de requêtes. Une évaluation expérimentale est présentée au niveau de la section 6. Une conclusion ainsi que des perspectives font l'objet de la section 7.

2. Travaux reliés à l'expansion de requêtes en RI

La problématique d'expansion de requêtes a été largement abordée par la communauté RI depuis deux décennies (Ruthven, 2003 ; Joho *et al.*, 2004 ; Kumaran et Allan, 2008). Les différentes approches proposées peuvent être groupées en deux principales classes selon le type de connaissances utilisé lors de l'expansion, résumées dans ce qui suit.

2.1. *Expansion à partir des collections de documents*

Ces approches d'expansion utilisent des connaissances dérivées à partir des collections de textes. Elles observent généralement la régularité des termes dans un contexte déterminé d'une collection de textes. Elles sont basées sur l'hypothèse qui stipule que "*L'emploi de deux termes en co-occurrence est l'expression d'une relation sémantique entre eux*" (Rijsbergen, 1979). L'avantage de ces approches est qu'elles sont faciles à mettre en œuvre tout en étant indépendantes du corpus.

Parmi les premiers travaux relatifs à cette classe d'approches, Grefensette (Grefenstette, 1992) contribue par une approche syntaxique pour l'extraction de contextes de mots à partir des corpus textuels. Le but de cette approche est de produire la liste des mots reliés à n'importe quel mot du corpus. Ces mots reliés seront utilisés dans l'expansion de requêtes.

D'autres approches s'appuient sur une analyse statistique des collections par l'extraction de règles d'association entre termes, afin d'ajouter des termes voisins à la requête originelle (Tangpong et Rungsawang, 2000 ; Haddad *et al.*, 2000 ; Latiri *et al.*, 2012). Les associations sont généralement basées sur la co-occurrence des termes dans les documents (Sun *et al.*, 2006 ; Lin *et al.*, 2008). L'usage d'une telle technique a montré que les liens inter-termes renforcent la notion de pertinence des documents par rapport aux requêtes. Par ailleurs, la principale limitation des approches d'expansion de requêtes par des règles d'association entre termes demeure le nombre très élevé d'associations découvertes, dont une large partie est redondante.

Une autre technique, répandue en RI, est la reformulation de requêtes par réinjection de pertinence, communément appelée *Relevance Feedback* (Ruthven et Lalmas, 2003). L'expansion de la requête se fait par de nouveaux termes issus de documents pertinents. Elle consiste à extraire, à partir d'un échantillon de documents jugés pertinents par l'utilisateur, les mots clés les plus pertinents et à les ajouter à la requête (Hlaoua *et al.*, 2010).

Une approche alternative, connue sous le nom de pseudo-réinjection de pertinence (Torjmen *et al.*, 2007), utilise des techniques de réinjection automatique de pertinence, en enrichissant la requête par les termes provenant des k premiers documents pertinents trouvés sans aucune intervention de l'utilisateur. En effet, la technique de réinjection automatique peut être bénéfique que si les requêtes initiales permettent de retrouver des documents pertinents ; dans le cas contraire, elle provoque une dégradation des performances. Une autre limite de cette approche est qu'elle utilise une technique d'expansion locale de requêtes, basée sur un ensemble de documents récupérés pour une requête donnée. Dans (Xu et Croft, 1996), les auteurs ont montré que l'utilisation de techniques d'expansion, basées sur une analyse globale, est plus efficace qu'une réinjection de pertinence locale.

2.2. *Expansion à base de ressources externes existantes*

Certaines approches d'expansion de requêtes utilisent un vocabulaire contrôlé issu de ressources lexicales et sémantiques externes, tels que les thésaurus (Hu *et al.*, 2009) et les ontologies (Song *et al.*, 2007). Un panorama d'approches d'expansion de requêtes à base de ressources externes de connaissances est présenté dans (Bhogal *et al.*, 2007).

Les premiers travaux, qui ont fait appel à des ressources lexicales externes de type dictionnaire ou thésaurus, ont d'abord tenté d'utiliser la base lexicale WordNet pour réaliser l'expansion de requêtes (Voorhees, 1993). Plus récemment, d'autres travaux ont proposé des approches d'enrichissement de requêtes utilisant Wikipédia comme collection externe (Koolen et Kamps, 2011).

En outre, parmi les approches utilisant les connaissances du domaine pour l'expansion de requêtes, les auteurs dans (Yang et Yao, 2005) ont proposé des mécanismes d'expansion de requêtes par des termes liés dans un réseau sémantique, représentant une base de connaissances spécifique d'un domaine. Dans (Revuri *et al.*, 2006), Revuri *et al.* définissent différents cas d'expansion de requêtes avec divers éléments d'une ontologie de domaine tels que les concepts, les propriétés et les instances. Une autre technique d'expansion sémantique de requêtes combinant les règles d'association entre termes et une ontologie de domaine est proposée dans (Song *et al.*, 2007).

En se positionnant par rapport aux approches d'expansion de requêtes de l'état de l'art, nous proposons, dans ce qui suit, une approche statistique d'expansion automatique de requêtes, basée sur les co-occurrences de termes. Les corrélations entre les termes sont extraites par une technique de fouille de textes globale de la collection de documents. Cette technique permet de générer des règles d'association entre termes dont la qualité est appréciée par la mesure de confiance et la mesure de l'intensité d'implication entropique.

3. Règles d'association entre termes et mesures de qualité

3.1. *Rappel sur les règles d'association entre termes*

En *Fouille de textes* (FT), une des principales méthodes produisant des connaissances sous forme de règles est l'extraction de règles d'association, introduite par (Agrawal et Skirant, 1994) .

Définition 1 *Une règle d'association entre termes, notée par RA, est une implication de la forme :* $R : T_1 \Rightarrow T_2$ *telles que* T_1 *et* T_2 *sont deux itemsets fréquents.* (Latiri *et al.*, 2010).

Effectivement, les règles d'association sont des tendances implicatives prémisse $\rightarrow$ conclusion où la prémisse et la conclusion sont des itemsets fréquents de la collection de textes. Une règle d'association entre termes est également appréciée par les deux métriques de *support* et de *confiance* (Agrawal et Skirant, 1994).

Le support de la règle d'association $R : T_1 \rightarrow T_2$ exprime, dans notre contexte de recherche, la fréquence avec laquelle deux termes T_1 et T_2 co-occurrent dans le corpus de textes. Autrement dit, la cardinalité de l'ensemble de phrases du corpus contenant en même temps les deux termes T_1 et T_2. La confiance de R exprime la probabilité conditionnelle pour qu'une phrase contienne T_2, sachant qu'elle contient le terme T_1.

Une règle d'association est dite *valide* si sa confiance est supérieure ou égale au seuil minimal de confiance $minconf$.

3.2. *La mesure de l'intensité d'implication entropique*

La confiance d'une règle d'association est considérée comme une mesure de co-occurrence qui se calcule simplement. Cependant, il est maintenant bien connu que la confiance présente trois inconvénients majeurs (Fleury *et al.*, 1995) : Elle varie linéairement. En outre, c'est un indice fréquentiel donc elle est insensible à la dilatation des effectifs. De plus, elle ne permet pas de rejeter les règles dues au hasard ou les règles évidentes.

Nous voulons introduire la notion d'entropie qui est très intéressante pour le traitement des textes. *L'intensité d'implication* est un indice de quasi-implication développé par Gras (Gras, 1996) et qui est au fondement d'une méthode d'analyse exploratoire de données nommée Analyse Statistique Implicative (ASI) [1] (Gras *et al.*, 2001). Cet indice est peu discriminant quand les cardinaux étudiés sont grands, comme toutes les

1. Une manière simplifiée de situer l'objet de l'Analyse Statistique Implicative (ASI), est de la comprendre comme « un champ théorique centré sur le concept d'implication statistique ou plus précisément sur le concept de quasi-implication pour le distinguer de celui d'implication logique des domaines de la logique et des mathématiques. L'étude de ce concept de quasi-implication en tant qu'objet mathématique, dans les champs des probabilités et de la statistique,

mesures de significativité statistique (Blanchard *et al.*, 2005). Pour résoudre ce problème, (Gras *et al.*, 2001) ont proposé de moduler les valeurs de l'intensité d'implication par un indice de quasi-implication descriptif fondé sur l'entropie de Shannon : *l'indice d'inclusion*. Le nouvel indice ainsi formé s'appelle *intensité d'implication entropique (IIE)*. De ce fait, (Blanchard *et al.*, 2004) donnent une définition plus générale que celle présentée dans (Gras *et al.*, 2001) de la mesure de *IIE* issue de la combinaison de l'intensité d'implication et de l'indice d'inclusion. Cette mesure est de nature statistique (grâce à l'intensité d'implication) tout en restant discriminante quand les cardinaux étudiés sont grands (grâce à l'indice d'inclusion). L'association des deux mesures est réalisée par la moyenne géométrique (Gras *et al.*, 2001). Ainsi, la *IIE* a été construite de façon à mieux mesurer la notion de qualité en intégrant à la fois l'étonnement statistique et la qualité inclusive.

Or, il est apparu que cette association présentait un caractère quelque peu artificiel, ce qui amène (Gras et Couturier, 2012) à une révision de cette formalisation. Cette révision conduit à une nouvelle *intensité d'implication entropique* associant cette fois l'*indice d'inclusion* i(a,b) et un *coefficient statistique n*. En effet, pour rendre l'indice d'inclusion sensible aux effectifs, ils ont introduit ce coefficient n, qui devra valoriser les observations sur de grandes jeux de données sur le plan statistique.

Définition 2 *L'intensité d'implication entropique d'une règle $a \rightarrow b$ est définie par :*

$$\Psi(a \rightarrow b) = (1 - \frac{1}{2\sqrt{n}}) i(a,b) \textit{ (Gras et Couturier, 2012)}$$

Pour extraire les règles avec *IIE*, nous reprenons l'algorithme Apriori améloiré (Gras et Couturier, 2012) permettant d'intégrer des calculs différents qui tiennent compte de l'entropie et en particulier de la nouvelle formule de l'intensité d'implication entropique.

3.3. *Dérivation des règles d'association entre termes appréciées par l'intensité d'implication entropique*

Pour effectuer l'expansion de requêtes d'une manière efficace, nous affirmons que la synergie entre les techniques de RI classiques et les méthodes de fouille de données avancées, en particulier les règles d'association (Agrawal et Skirant, 1994), est particulièrement appropriée selon des études antérieures (Haddad *et al.*, 2000 ; Tangpong et Rungsawang, 2000 ; Lin *et al.*, 2008 ; Latiri *et al.*, 2012). En effet, le paradigme d'extraction de règles d'association est l'extraction de motifs fréquents co-occurant dans une base de données de transaction. Dans notre cas, les motifs extraits sont des termes simples.

---

a permis de construire des outils théoriques qui instrumentent une méthode d'analyse de données. (Régnier *et al.*, 2009).

Il est intéressant de mentionner que le processus de génération des règles d'association, à partir d'une collection de documents, est en effet réalisé en amont. Nous utilisons l'algorithme *Apriori* pour l'extraction d'une base de règles d'association appréciées par la *IIE*, noté par BRIIE. Nous déployons une version améliorée (Gras et Couturier, 2012) d'*Apriori* permettant la dérivation des règles d'association. Afin de générer ces règles, nous avons fait varier les seuils minimaux de support, *i.e.*, *min-supp*. Rappelons que ces seuils sont définis pour éliminer, respectivement, les règles très rares et celles qui sont très fréquentes. Il importe de préciser que les valeurs de ces seuils ont été déterminées empiriquement, en étudiant la régularité sur la fréquence d'apparition des termes respectives aux corpus des différentes collections utilisées dans une même langue. Les règles d'associations monolingues extraites sont de la forme : $hyp \rightarrow conc$ sont appréciées par les mesures de support, confiance et de l'intensité d'implication entropique *(IIE)*. Avant d'entamer le processus d'expansion automatique de requêtes, il est nécessaire de sélectionner parmi la base de règles d'association extraites celles ayant les valeurs des mesures de confiance et/ou de l'*IIE* supérieure ou égale à leurs seuils minimaux.

4. Construction des méta-règles de contexte

Le processus de fouille de textes s'appuyant sur l'extraction des règles d'association à l'aide d'un algorithme engendre une quantité importante de règles d'association. Ces règles extraites nécessitent une étape de post-traitement pour disposer d'une base de règles réduite en se basant sur des mesures de qualité. Nous proposons une nouvelle structure de données afin de réduire cette base de règles. Ce processus consiste à regrouper les règles en classes selon des critères fiables.

Notre but est de regrouper les règles en classes dont l'étiquette est la prémisse. Ces nouvelles règles offrent des informations sur les corrélations de termes inter-documents. De ce fait, ces règles permettent d'expliciter des relations plus fines entre des termes qui apparaissent ensemble. Ce qui nous amène à une nouvelle structure de données *les méta-règles de contexte (MR)*. De cette façon, chaque *MR* regroupe une classe de règles ayant en commun la même prémisse et sélectionnées auparavant sur la base de mesures de qualité. La *MR* est de la forme $A \rightarrow B$ où A est la prémisse commune à un ensemble de règles et B la conclusion regroupant les parties conclusion de ces mêmes règles. Le problème qui se pose est la manière de construire cette nouvelle structure de données avec les mesures de qualité correspondantes, en particulier le support, la confiance et la IIE. Plusieurs situations peuvent se présenter pour calculer les mesures de qualité pour chaque méta-règle de contexte. Ces mesures de qualité se calculent en se basant sur les mesures des règles d'association utilisées dans la construction d'une méta-règle :

– Choisir pour chaque mesure la valeur maximale parmi les valeurs correspondantes aux règles d'association utilisées dans la construction du méta-règle, notée *max*.

– Choisir pour chaque mesure la valeur minimale parmi les valeurs correspondantes aux règles d'association utilisées dans la construction du méta-règle, notée *min*.

– Calculer une qualité moyenne entre les différentes valeurs correspondantes à la mesure pour les règles d'association utilisées. La moyenne géométrique est généralement utilisée pour une combinaison d'indices. Nous choisissons de calculer alors la moyenne géométrique des valeurs de chaque mesure pour les éléments de la partie conclusion d'une méta-règle, notée *moy*.

La formule pour calculer la mesure de qualité MQ d'une MR selon la moyenne géométrique est définie par :

$$MQ = (x_1 \times \mathrm{x}_2 \times \ldots \times \mathrm{x}_i)^{1/n}$$

x_i : la valeur du *i*ème élément du MR.
n : le nombre des éléments de la partie conclusion du MR.

Exemple 1 *Soient les règles d'association de la forme A → B (support ; confiance) suivantes ayant en commun la même prémisse "yesterday" :*
yesterday → years (36.23 ; 32.78)
yesterday → time (36,23 ; 30,50)
yesterday → people (36,23 ; 25,28)
yesterday → Glasgow (36,23 ; 16,55)
⇒ La MR résultante est : yesterday → years time people Glasgow (36.23 ; 25.43)
tels que :
$MQ1$ = $(36,23 \times$ *36,23* $\times$ *36,23* $\times$ *36,23* $)^{1/4}$ = *36,23*
$MQ2$ = $(32,78 \times$ *30,50* $\times$ *25,28* $\times$ *16,55* $)^{1/4}$ = *25,43*

5. Processus d'expansion automatique de requêtes par les règles d'association de la base BRIIE

5.1. *Expansion automatique des requêtes par les règles d'association de la base BRIIE*

Notre approche d'expansion automatique de requêtes consiste à dériver, dans un premier temps, la base BRIIE de règles d'association non-redondantes entre termes à partir d'une collection de documents, et de l'utiliser, dans un deuxième temps, pour étendre la requête originelle de l'utilisateur. Tous les champs de la requête, *i.e.,* titre, champs descriptifs et narratifs, sont utilisés lors du processus d'expansion. Il importe de souligner que la base BRIIE est mieux adaptée au processus d'expansion automatique de requêtes, dans lesquelles l'ensemble des termes originels sera étendu par les conclusions des règles valides de la base BRIIE, et ayant les termes de la requête initiale dans leurs prémisses respectives. Rappelons que les règles d'association entre termes de la base BRIIE sont non-redondantes et qu'elles sont dotées d'une prémisse minimale et une conclusion maximale, ce qui offre plus de termes candidats pour l'expansion.

À partir des règles d'association entre termes de la base BRIIE, chaque terme de la requête est traité individuellement en le cherchant dans les prémisses minimales des règles valides. La requête sera ensuite enrichie par la conclusion maximale de chaque règle ayant comme prémisse le ou les termes de la requête originelle. De part sa forme maximale, la conclusion de la règle offre plus de termes candidats pour l'expansion.

5.2. *Expansion automatique des requêtes par les méta-règles de contexte*

La même approche d'expansion automatique de requêtes est appliquée pour les méta-règles de contexte. Il s'agit d'étendre chaque requête originelle de la collection par tous les termes qui apparaissent dans la partie conclusion des méta-règles de contexte, qui ont dans leurs prémisses respectives les termes de la requête originelle. Tous les champs de la requête, *i.e.,* titre, champs descriptifs et narratifs, sont utilisés lors du processus d'expansion. Notre approche d'expansion automatique de requêtes consiste à construire, dans un premier temps, les méta-règles à partir de la base BRIIE, et de l'utiliser, dans un deuxième temps, pour étendre la requête originelle de l'utilisateur. L'ensemble des termes originels sera étendu par la partie conclusion des méta-règles, et ayant les termes de la requête initiale dans leurs prémisses respectives.

Nous sélectionnons parmi l'ensemble des méta-règles celles dont leurs nouvelles mesures de confiance et IIE sont supérieures ou égale à un seuil minimal fixé pour chaque mesure.

6. Évaluation expérimentale de l'approche d'expansion

6.1. *Cadre d'évaluation*

L'évaluation expérimentale a été menée sur la plateforme TERRIER[2] considérée comme un moteur d'indexation et de recherche pendant toutes les expérimentations décrites ci-dessous. Nous utilisons la collection de test décrite dans le tableau 1. Nous avons testé notre approche d'expansion sur le schéma de pondérations OKAPI BM25 qui est une méthode d'ordonnancement de la méthode OKAPI, la plus connue des méthodes probabilistes, et ayant pour but de construire un modèle probabiliste qui prend en compte la fréquence des termes ainsi que la taille des documents (Jones *et al.*, 2000).

6.2. *Collection de test*

Nous utilisons, pour tester notre approche, la collection de textes en anglais du corpus CLEF 2003. Quelques caractéristiques statistiques de la collection en anglais

2. *http ://terrier.org*

The Glasgow Herald 1995, notée dans la suite de l'article *GH-95*, sont données dans le tableau 1. Nous avons mené les tests sur 60 requêtes, chacune étant fournie avec un ensemble de documents jugés pertinents pour l'anglais.

Collection	GH-95
Nombre de documents	56472
Nombre de phrases	1149630
Taille du vocabulaire	12513000
Nombre de requêtes	60

Tableau 1. *Caractéristiques de la collection* THE GLASGOW HERALD 1995 (GH-95) *de CLEF 2003.*

Le corpus a été pré-traité de façon à normaliser les mots qu'il contient. Premièrement, le corpus est nettoyé et étiqueté afin d'extraire les substantifs (les noms et les adjectifs). Seuls les noms et les adjectifs sont pris en compte par nos algorithmes de génération des RAs.

6.3. *Résultats et discussions*

Nous évaluons nos scénarios pour la recherche monolingue avec des requêtes en anglais pour interroger le corpus en anglais de la collection *GH-95*. Les résultats de RI monolingue sont reportés dans le tableau 2. Dans ce tableau, "*QE-conf*>= .. %" fait référence à l'expansion de requêtes avec les règles dont la confiance est supérieure ou égale à .. et "*IIE*>= .. %" fait référence à l'expansion de requêtes avec les règles dont la IIE est supérieure ou égale à .. .

Les requêtes étendues sont évaluées selon les mêmes mesures que les requêtes originelles. Les résultats sont comparés avec ceux obtenus sans expansion (*Baseline*) pour calculer l'amélioration effectuée par l'injection des termes à partir des règles d'association entre termes. Il est à noter que la variation de la pertinence système, dénoté par Δ est calculée comme suit (Latiri *et al.*, 2012) :

$$\Delta = \frac{(Result\ with\ expansion) - (Result\ without\ expansion)}{(Result\ without\ expansion)} \quad [1]$$

6.3.1. *Évaluation des règles d'association entre termes appréciées par la IIE*

Le tableau 2 présente, pour le corpus *GH-95*, l'évolution de la précision moyenne (MAP) et de la précision moyenne à 11 points de rappel (11PT_AVG) pour :

– Scénario 1 *QE-conf* : qui consiste à étendre les requêtes en utilisant les règles de la base BIIE, sélectionnées selon la mesure de CONFIANCE.

Evaluation	MAP		11pt_avg	
Baseline	0.2798		0.2890	
QE-conf	QE	Δ	QE	Δ
>=40%	0.3031	+8.33%	0.3113	+7.72%
>=50%	0.3066	+9.58%	0.3153	+9.1%
>=60%	0.3079	+10.04%	0.3161	+9.38%
>=70%	0.3079	+10.04%	0.3161	+9.38%
QE-IIE	QE	Δ	QE	Δ
>=50%	0.3056	+9.22%	0.3140	+8.65%
>=60%	0.3057	+9.26%	0.3140	+8.65%
>=70%	0.3066	+9.58%	0.3153	+9.1%
>=80%	0.3086	+10.29%	0.3170	+9.69%
QE-conf_IIE	QE	Δ	QE	Δ
50%_50%	0.3066	+9.58%	0.3153	+9.1%
50%_60%	0.3066	+9.58%	0.3153	+9.1%
50%_70%	0.3066	+9.58%	0.3153	+9.1%
50%_80%	0.3086	+10.29%	0.3170	+9.69%

Tableau 2. *Apport des règles d'association appréciées par la IIE (minsupp=7%) dans l'expansion de requêtes en anglais de la collection GH-95 en MAP et 11pt_avg avec* Δ *en %.*

– Scénario 2 *QE-IIE* : qui consiste à étendre les requêtes en déployant les règles de la base BRIIE appréciées par la IIE.

– Scénario 3 *QE-conf_IIE* : qui consiste à étendre les requêtes en utilisant les règles de la base BRIIE, sélectionnées selon la mesure de la CONFIANCE et de la IIE simultanément.

Ces scénarios sont comparés à la base comparative (*Baseline*).

Les stratégies d'expansion effectuées améliorent la *11pt_avg* et la *MAP*. En effet, nous observons une amélioration progressive en *MAP* et en *11pt_avg* en augmentant la valeur de la confiance et de la *IIE*. Le Scénario 1 *(QE-conf)* présente une amélioration en *MAP* et en *11pt_avg* pour toutes les stratégies effectuées en faisant varier la valeur de la confiance, comparé au *Baseline*. Cependant, nous observons en même temps, pour le scénario *QE-conf*, une stabilité de la valeur de la *MAP* de 0.3079 et de la *11pt_avg* de 0.3161 entre les scénarios *QE-conf>=60%* et *QE-conf>=70%*. Ceci est du à l'utilisation du même ensemble de règles pour l'expansion. Nous notons ainsi, pour le Scénario 2 *(QE-IIE)*, que les règles utilisées pour l'expansion de requêtes et sélectionnées selon la *IIE* induisent une amélioration en moyenne meilleure que celle en utilisant les règles sélectionnées par la mesure de confiance. Cette amélioration se présente par la valeur de la *IIE* la plus importante *(IIE>=80%)* avec une amélioration en *MAP* de 10.29% et en *11pt_avg* de 9.69%. Pour le Scénario 3, nous voulons visualiser l'impact de la combinaison des deux mesures de confiance et de *IIE*. Selon les résultats

du Scénario 1 et 2, nous choisissons une valeur de confiance moyenne (conf>=50%) en augmentant les valeurs de la *IIE*. Ainsi, nous constatons une amélioration de ces scénarios par rapport au *Baseline*. De plus, nous remarquons une stabilité de la *MAP* (0.3066) et de la *11pt_avg* (0.3153) pour les scénarios *QE-conf>=50%_IIE>=50%*, *QE-conf>=50%_IIE>=60%* et *QE-conf>=50%_IIE>=70%*. Ceci est du au déploiement du même ensemble des règles pour ces scénarios. Autrement dit, la plupart des règles dont la confiance est supérieure ou égale à 50% possèdent une *IIE* supérieure ou égale à 70%. Selon les scénarios utilisant les règles avec *IIE*, la meilleure *MAP* et *11pt_avg* sont données par les règles dont la *IIE* est supérieure ou égale à 80%, avec une amélioration de 10.29% pour la *MAP* et de 9.69% pour la *11pt_avg* pour les scénarios *(IIE>=80%)* et *(conf>=50%_IIE>=80%)*. Par conséquent, nous constatons que la pertinence système atteint sa meilleure amélioration avec une valeur de la *IIE* relativement élevée (en général à partir de 50%).

6.3.2. *Évaluation des méta-règles de contexte*

Nous construisons les méta-règles en fixant des valeurs moyennes pour la *confiance* (= 50%) et la *IIE* (= 50%). Notre but principal est de construire des méta-règles à partir de règles intéressantes avec un nombre acceptable de règles. En plus, le choix d'une valeur faible de *confiance* ou de *IIE* permet de construire des méta-règles bruitées, ainsi que le choix d'une valeur élevée de *confiance* ou de *IIE* sélectionne un nombre faible de règles pour construire les méta-règles. Nous voulons alors interpréter l'apport des méta-règles par rapport aux règles d'association dans l'expansion de requêtes.

D'une part, nous construisons les méta-règles avec les règles d'association dont la confiance est >= 50% (noté par *conf>=50%_MR*). Nous voulons effectuer un filtrage de ces méta-règles avec la nouvelle IIE calculée pour les utiliser en expansion de requêtes. Nous interprétons auparavant (tableau 2) que la plupart des règles dont la confiance est supérieure ou égale à 50% possède une IIE supérieure ou égale à 70%. Nous observons la même interprétation pour la nouvelle IIE des méta-règles. Pour cette raison, nous sélectionnons les méta-règles (conf>=50%_MR) selon IIE>=70% (noté par *conf>=50%_MR_IIE>=70%*) et IIE>=80% (noté par *conf>=50%_MR_IIE>=70%*).

D'autre part, nous construisons les méta-règles avec les règles d'association dont la IIE est >= 50% (noté par *IIE>=50%_MR*). Nous voulons effectuer un filtrage de ces méta-règles avec la nouvelle confiance calculée pour les utiliser en expansion de requêtes. Nous observons à partir de ces méta-règles, que la nouvelle confiance ne dépasse pas 60%. Pour cette raison, nous sélectionnons les méta-règles (IIE>=50%_MR) selon conf>=50% (noté par *IIE>=50%_MR_conf>=50%*) et conf>=40% (noté par *IIE>=50%_MR_conf>=40%*).

Nous évaluons chaque scénario de méta-règles en précisant la stratégie de calcul des nouvelles mesures de *confiance* et de *IIE* :

– Choisir pour chaque mesure la valeur maximale parmi les valeurs correspondantes aux règles d'association utilisées dans la construction de la méta-règle, notée *max*,

– Choisir pour chaque mesure la valeur minimale parmi les valeurs correspondantes aux règles d'association utilisées dans la construction de la méta-règle, notée *min*,

– Calculer la moyenne géométrique entre les différentes valeurs correspondantes à la mesure pour les règles d'association utilisées, notée *moy*.

Evaluation		MAP		11pt_avg	
conf>=50%		0.3066		0.3153	
conf>=50%_MR		QE	Δ	QE	Δ
_IIE>=70	max	0.3066	0	0.3153	0
	min	0.3066	0	0.3153	0
	moy	0.3066	0	0.3153	0
_IIE>=80	max	0.3058	-0.26%	0.3143	-0.32%
	min	0.3083	+0.55%	0.3167	+0.44%
	moy	0.3060	-0.2%	0.3143	-0.32%
IIE>=50%		0.3056		0.3140	
IIE>=50%_MR		QE	Δ	QE	Δ
_conf>=40%	max	0.3056	0	0.3140	0
	min	0.3051	-0.16%	0.3133	-0.22%
	moy	0.3056	0	0.3140	0
_conf>=50%	max	0.3056	0	0.3140	0
	min	0.3077	+0.69%	0.3159	+0.6%
	moy	0.3073	-0.56%	0.3157	-0.54%

Tableau 3. *Apport des méta-règles par rapport aux règles d'association dans l'expansion de requêtes en anglais de la collection GH-95 en MAP et 11pt_avg avec* Δ *en %.*

Les résultats obtenus avec les méta-règles sont comparés avec ceux obtenus avec les règles d'association. Á la lecture du tableau 3, nous constatons que le déploiement des méta-règles réalise des améliorations minimes dans l'expansion de requêtes par rapport aux règles d'association. Nous distinguons des améliorations minimes pour les scénarios *conf>=50%_MR_IIE>=80%* et *IIE>=50%_MR_conf>=50%* utilisant la stratégie *min* pour calculer les nouvelles mesures de *confiance* et de *IIE* pour les méta-règles construites. Il peut s'avérer alors judicieux de choisir le minimum entre les différentes valeurs correspondantes à un élément de la partie conclusion d'une méta-règle pour déterminer les nouvelles valeurs de chaque mesure.

7. Conclusion et travaux en cours

L'objectif est de sélectionner les règles afin de les déployer dans l'expansion automatique de requêtes. Cette sélection permet de réduire le nombre de règles à analyser. Nous sommes intéressés par l'extraction de règles d'association non redondantes entre les termes avec la *IIE*. Nous avons comparé l'expansion de requêtes avec ces règles avec celles choisies en fonction de la confiance. Nous concluons que les règles sélectionnées en fonction de la *IIE* sont meilleures que celles sélectionnées par la confiance. Cette mesure fournit des améliorations intéressantes en RI pour l'expansion de la requête. D'une autre coté, nous construisons les méta-règles à partir de cette base des règles d'association. Les nouvelles mesures de chaque méta-règles sont calculées selon différentes stratégies. Nous sélectionnons ces méta-règles selon les nouvelles mesures. Nous constatons, que la stratégie consistant à choisir la valeur minimale entre les différentes valeurs de chaque mesure, donne des améliorations minimes par rapport aux résultats obtenus par les règles d'association.

Notre travail se déroule dans le contexte de la RI monolingue en déployant des règles d'association monolingues et construire à partir desquelles les méta-règles dans la même langue. Notre perspective est de continuer ce travail en RI multilingue en utilisant les méta-règles pour la construction d'un lexique bilingue.

8. Remerciement

Ce travail est partiellement financé par le projet franco-tunisien PHC Utique nř 14G 1404, intitulé RIMS-FD.

9. Bibliographie

Agrawal R., Skirant R., « Fast algorithms for Mining Association Rules », *Proceedings of the 20th International Conference on Very Large Databases VLDB 1994*, Santiago, Chile, p. 478-499, September, 1994.

Bhogal J., Macfarlane A., Smith P., « A review of ontology based query expansion », *Information Processing and Management*, vol. 43, nº 4, p. 866 - 886, 2007.

Blanchard J., Guillet F., Briand, Gras R., « Ipee : Indice probabiliste d'Ãl'cart Ãă l'Ãl'quilibre pour l'Ãl'valuation de la qualitÃl' des rÃĺgles », *RNTI*, vol. E-5, p. 391-395, 2005.

Blanchard J., Kuntz P., Guillet F., Gras R., « Mesure de qualité des règles d'association par l'intensité d'implication entropique », *Revue Nationale des Technologies de l'Information, RNTI*, 2004.

Buckley C., Salton G., Allan J., Singhal A., « Automatic Query Expansion Using SMART : TREC-3 », *Proceedings of the* 3^{rd} *Text REtrieval Conference*, 1994.

Fleury L., Djeraba C., Philippe J., « Rule evaluation for knowledg discovery in databases », *in* N. Revell, A. M. Tjoa (eds), *Proceedings of the 6th Conference on Database and Expert System Applications DEXA'95,*, vol. 978 of *Lecture Notes in Computer Science*, London, United Kingdom, p. 405-414, September, 1995.

Gras R., *L'implication statistique : Nouvelle méthode exploratoire de données*, la pensée sauvage editions edn, 1996.

Gras R., Couturier R., *L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire*, IUFM de l'Université de Caen Basse Normandie, chapter Implication entropique et causalité, p. 39-50, 2012.

Gras R., Kuntz P., Couturier R., Guillet F., « Une version entropique de l'intensité d'implication pour les corpus volumineux. », *in* H. Briand, F. Guillet (eds), *EGC*, vol. 1 of *Extraction des Connaissances et Apprentissage*, Hermes Science Publications, p. 69-80, 2001.

Grefenstette G., « Use of semantic context to produce term association lists for text retrieval », *Proceedings of the* 15^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'92*, ACM Press, Copenhagen, Denmark, p. 89-97, June, 1992.

Haddad H., Chevallet J. P., Bruandet M. F., « Relations between Terms Discovered by Association Rules », *Proceedings of the Workshop on Machine Learning and Textual Information Access in conjunction with the* 4^{th} *European Conference on Principles and Practices of Knowledge Discovery in Databases, PKDD 2000*, Lyon, France, September, 2000.

Hlaoua L., Pinel-Sauvagnat K., Boughanem M., « Relevance feedback revisited : dealing with content and structure in XML documents », *Int. J. on Digital Libraries*, vol. 11, n° 1, p. 1-24, 2010.

Hu J., Wang G., Lochovsky F. H., Sun J.-T., Chen Z., « Understanding user's query intent with Wikipedia », *Proceedings of the* 18^{th} *International Conference on World Wide Web, WWW 2009*, ACM Press, Madrid, Spain, p. 471-480, April, 2009.

Joho H., Sanderson M., Beaulieu M., « A Study of User Interaction with a Concept-Based Interactive Query Expansion Support Tool », *Proceedings of the 26th European Conference on Information Retrieval Research, ECIR 2004*, vol. 2997 of *LNCS*, Springer-Verlag, Sunderland, UK, p. 42-56, April, 2004.

Jones K. S., Walker S., Robertson S. E., « A probabilistic model of information retrieval : development and comparative experiments », *Information Processing and Management*, vol. 36, n° 6, p. 779-840, 2000.

Koolen M., Kamps J., « Are Semantically Related Links More Effective for Retrieval », *Proceedings of the* 33^{rd} *European Conference on IR Research, ECIR 2011*, vol. 6611 of *LNCS*, Springer-Verlag, Dublin, Ireland, p. 92-103, April, 2011.

Kumaran G., Allan J., « Adapting information retrieval systems to user queries », *Information Processing and Management*, vol. 44, n° 6, p. 1838-1862, 2008.

Latiri C., Haddad H., Hamrouni T., « Towards An Effective Automatic Query Expansion Process Using An Association Rule Mining Approach », *Journal of Intelligent Information Systems*, vol. 39, n° 1, p. 209-247, 2012.

Latiri C., Slimani Y., Nasri C., Smaïli K., « Extraction des séquences fermées fréquentes à partir de corpus parallèles : application à la traduction automatique », *Actes des dixièmes journées francophones en Extraction et gestion des connaissances, EGC'2010*, vol. RNTI-E-19 of *Revue des Nouvelles Technologies de l'Information*, Cépaduès-Éditions, Hammamet, Tunisie, p. 55-60, 2010.

Lin H. C., Wang L. H., Chen S. M., « Query Expansion for Document Retrieval by Mining Additional Query Terms », *Information and Management Sciences*, vol. 19, n° 1, p. 17-30, 2008.

Régnier J.-C., Gras R., Guillet F. (eds), *Analyse Statistique Implicative. Une méthode d'analyse de données pour la recherche de causalités.*, cépaduès ÃL'ditions edn, RNTI-E-16, Toulouse, 2009.

Revuri S., Upadhyaya S. R., Kumar P. S., « Using Domain Ontologies for Efficient Information Retrieval », *Proceedings of the* 13^{th} *International Conference on Management of Data*, Tata McGraw-Hill Publishing, Delhi, India, p. 170-173, December, 2006.

Rijsbergen C. V., *Information Retrieval*, Butterworths, London, 1979.

Rungsawang A., Tangpong A., Laohawee P., Khampachua T., « Novel Query Expansion Technique Using Apriori Algorithm », *Proceedings of the* 8^{th} *Text REtrieval Conference, TREC 8*, Gaithersburg, Maryland, p. 453-456, November, 1999.

Ruthven I., « Re-examining the potential effectiveness of interactive query expansion », *Proceedings of the* 26^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2003*, ACM Press, Toronto, Canada, p. 213-220, July/August, 2003.

Ruthven I., Lalmas M., « A survey on the use of relevance feedback for information access systems », *Knowledge Engineering Review*, vol. 18, n° 2, p. 95-145, 2003.

Salton G., McGill M. J., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

Song M., Song I., Hu X., Allen R. B., « Integration of association rules and ontologies for semantic query expansion », *Data and Knowledge Engineering*, vol. 63, n° 1, p. 63 - 75, 2007.

Sun R., Ong C., Chua T., « Mining dependency relations for query expansion in passage retrieval », *Proceedings of the* 29^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2006*, ACM Press, Seattle, Washington, USA, p. 382-389, August, 2006.

Tangpong A., Rungsawang A., « Applying Association Rules Discovery in Query Expansion Process », *Proceedings of the* 4^{th} *World Multi-Conference on Systemics, Cybernetics and Informatics, SCI 2000*, Orlando, Florida, USA, July, 2000.

Torjmen M., Pinel-Sauvagnat K., Boughanem M., « Using Pseudo-Relevance Feedback to Improve Image Retrieval Results », *Proceedings of the* 8^{th} *Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, vol. 5152 of *LNCS*, Springer-Verlag, Budapest, Hungary, p. 665-673, September, 2007.

Voorhees E. M., « Using WordNet to Disambiguate Word Senses for Text Retrieval », *Proceedings of the* 16^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1993*, ACM Press, Pittsburgh, PA, USA, p. 171-180, June/July, 1993.

Xu J., Croft W. B., « Query Expansion Using Local and Global Document Analysis », *Proceedings of the* 19^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1996*, ACM Press, Zurich, Switzerland, p. 4-11, August, 1996.

Yang X., Yao Y., « Conceptual query expansion », *Proceedings of the Atlantic Web Intelligence Conference, AWIC 2005*, vol. 3528 of *LNCS*, Springer-Verlag, Lodz, Poland, p. 190-196, June, 2005.