Classification automatique de textes basée sur des hiérarchies de concepts

Kurt Englmeier (**), G. Hubert (*), Josiane Mothe (*, ***)
mothe@irit.fr, hubert@irit.fr, kurt@diwsysv.diw-berlin.de,

(*) Institut de Recherche en Informatique de Toulouse, 118 route de Narbonne, 31062 Toulouse CEDEX 04, France
(**) German Institute for Economic Research (DIW), Königin-Luise-Str. 5, 14195 Berlin, Germany
(***) Institut Universitaire de Formation des Maîtres, 56 Avenue de l'URSS, 31400 Toulouse, France

Mots-clés :
Recherche d'information, hiérarchies de concepts, Web sémantique, classification automatique
Keywords :
Information Retrieval, Concept hierarchies, semantic Web, Automatic clustering

Résumé

Cet article présente une méthode de classification automatique de textes à partir de hiérarchies de concepts décrivant un domaine. Cette méthode est basée sur un vote qu'obtiennent les concepts en fonction des termes qui peuvent être extraits des contenus de documents. Nous avons étudié l'influence de différents paramètres sur les résultats obtenus. L'évaluation ainsi que ses résultats sont présentés. Cette approche pourrait permettre d'offrir des mécanismes de consultation du Web via sa structuration autour de hiérarchies de concepts.

1 Introduction

Le Web constitue une des plus vastes collections d'information concernant à peu près tous les domaines. Un des majeurs problèmes à résoudre concerne l'accès efficace à cette information, que ce soit par des acteurs humains ou par des robots (agents).
Actuellement, lorsqu'un utilisateur recherche de l'information sur un domaine, il a généralement recours à un moteur de recherche. Ces moteurs utilisent différents moyens pour organiser et accéder à l'information. On distingue généralement deux techniques. La première technique utilisée par les moteurs consiste à maintenir des index qui associent des termes extraits des contenus des pages et les URL correspondantes. Un autre principe, utilisé par Yahoo! par exemple, consiste à classifier les URL en fonction de catégories hiérarchiques. Ce dernier principe est à rapprocher des techniques d'indexation manuelle basées sur un langage contrôlé ou sur un thésaurus (généralement hiérarchique). On peut citer par exemple le thesaurus MeSH de la National Library of Medecine qui est utilisé pour indexer les documents de MedLine. De notre point de vue, ce principe d'indexation ou de classification contrôlée se rapproche du projet de Web sémantique initié par le W3C. Ce projet a pour objectif de structurer le Web actuel afin de faciliter en particulier des traitements automatiques intelligents et de permettre une coopération plus étroite entre utilisateurs et agents. Ce projet doit reposer sur les technologies XML et RDF qui permettent d'ores et déjà d'ajouter des éléments de structure aux contenus des pages. Un autre élément de base du Web sémantique correspond aux ontologies [2]. Une ontologie, qui en philosophie réfère à la science de l'existence, correspond pour les acteurs du Web aux relations formelles entre termes. Une façon d'ajouter de la structure au Web consiste donc à créer des liens entre le contenu des pages et une ou plusieurs ontologies, apportant ainsi une sémantique utile pour les moteurs de recherche par exemple.
Notre approche vise à générer automatiquement des liens entre des pages Web et des ontologies matérialisées par des hiérarchies de concepts qui correspondent à la connaissance d'un domaine. Ce

lien est matérialisé au travers d'un code XML qui n'est pas interprété par les navigateurs mais qui pourrait être utilisé par des agents intelligents. Dans notre approche, les ontologies peuvent également être utilisées comme base pour naviguer et accéder à l'information utile.
Les travaux présentés dans cet article s'inscrivent dans le cadre du projet IRAIA [5]. IRAIA est un système d'accès à l'information permettant une recherche contextuelle dans de larges espaces de données notamment associés à l'économie. Avec IRAIA, les utilisateurs consultent les informations en sélectionnant des entrées (termes ou groupes de mots) dans des hiérarchies de concepts spécifiques. Ainsi, une requête est simplement définie en cliquant sur des entrées, des concepts signifiants. Chacune de ces hiérarchies de concepts correspond à un aspect ou à une catégorie. Par exemple, l'espace d'informations associé à "Economie" peut être groupé en catégories comme les régions, les industries et les variables ou indices économiques. Les HC structurent l'espace d'information et correspondent à la connaissance associée à des domaines. Dans ce type d'espace d'information l'utilisateur ne perd jamais le contexte sémantique lors de sa recherche et il ne peut pas rencontrer par hasard des données qui appartiennent à un autre contexte. Le fait de grouper les données en fonction d'espaces identifiés résout le problème d'ambiguïté sur lequel les moteurs de recherche traditionnels butent. Selon l'approche choisie, l'utilisateur cible l'information que lui retourne le système en choisissant les entrées dans les différentes hiérarchies de concepts relatives au domaine interrogé. Ces hiérarchies sont utilisées pour classifier les documents mais également servent de langage d'interrogation. Dans cet article, nous nous intéressons au premier aspect. L'interface d'interrogation quant à elle est présentée dans [3]. Dans la partie 2 de cet article, nous présentons la méthode de classification automatique des documents textuels selon des hiérarchies de concepts prédéfinies. Dans la partie 3, nous décrivons le cadre expérimental que nous avons utilisé pour évaluer les performances de notre méthode. Enfin la partie 4 présente les résultats que nous avons obtenus.

2 Association de textes à des hiérarchies de concepts

Une HC est une arborescence composée de concepts ou entrées, chaque entrée correspondant à un ensemble de termes.
Ainsi, l'association automatique des textes à des HC peut être vue comme la classification de documents suivant des domaines de connaissances ; elle peut également être vue comme l'indexation automatique de documents à partir d'un vocabulaire contrôlé issu des HC, l'ensemble des entrées auxquelles un texte est rattaché correspondant alors à l'annotation du texte. Cette association automatique permet de créer des contextes de recherche non ambigus et clairement identifiés.

2.1 Description générale de la méthode

Chaque texte peut ainsi être associé à différentes entrées d'une même hiérarchie ou de hiérarchies différentes. L'association d'un texte à une entrée d'une hiérarchie repose sur la méthode Vector Voting [8]. Cette méthode se base sur l'extraction automatique des termes de chacune des entrées dans le contenu du texte. L'importance de l'association du texte avec une entrée donnée est calculée par une méthode de vote, qui peut être rapprochée de la méthode HVV (Hyperlink Vector Voting) utilisée dans le contexte du Web pour calculer la pertinence d'une page en fonction des sites qui y référent [7]. Dans notre contexte, plus l'entrée ou une partie de l'entrée est présente dans le texte, plus le lien entre le texte et cette entrée sera fort.

2.2 Etapes de la méthode

Le principe d'association d'un document à des entrées s'effectue suivant différentes étapes :

- Extraction automatique des termes représentatifs de chaque entrée d'une hiérarchie de concepts et de leur importance dans l'entrée.
- Extraction automatique des termes représentatifs du document et de leur importance au sein du document. Le processus d'extraction est basé sur un ensemble de règles qui utilisent des balises des documents et des expressions régulières. Une fois les balises détectées, des

fonctions sémantiques et syntaxiques complètent le processus d'extraction afin de gérer les synonymes et l'élimination de termes inintéressants.

- Pour chaque entrée de la hiérarchie, calcul du score obtenu pour chaque entrée de la hiérarchie selon une méthode de vote. Le calcul de score peut être basé sur différentes fonctions de calcul qui peuvent faire intervenir des mesures comme l'importance d'un terme dans le document, l'importance d'un terme dans la hiérarchie, la taille du document, la taille de la hiérarchie, le nombre de termes d'une entrée présents dans le document.
- Classement des entrées de la hiérarchie dans l'ordre des scores obtenus, puis sélection de l'ensemble des entrées à associer au document suivant une stratégie définie (par exemple les entrées ayant obtenu un score supérieur à un seuil donné, ou les *n* premières entrées ayant les meilleurs scores).
- Modélisation des associations entre document et entrées de hiérarchies sous forme de code XML ajouté au contenu du document. Ce code bien que n'étant pas interprété par les navigateurs actuels peut être exploité par des agents intelligents.

3 Mise en œuvre et expérimentation

Le principe d'association entre documents et HC a été mis en œuvre dans le cadre du projet IRAIA [5]. L'influence de différents paramètres a été étudiée.

3.1 Paramètres évalués

Différentes fonctions de vote ont été évaluées. Elles mettent en œuvre les facteurs suivants :

(1) $\frac{F(T,D)}{S(D)}$ où F(T,D) correspond à l'occurrence du terme T dans le document D et S(D) correspond à la taille de D. Ce facteur mesure donc l'importance du terme T dans le document D.

(2) $\frac{S(H)}{F(T,H)}$ où F(T,H) correspond à l'occurrence du terme T dans la HC H et S(H) correspond à la taille de H. Ce facteur mesure donc l'importance du terme T dans H.

(3) $\frac{NT(E,D)}{NT(E)}$ où NT(E) correspond au nombre de termes de l'entrée E et NT(E,D) correspond au nombre de termes de l'entrée E qui apparaissent dans D. Ce facteur mesure donc le taux de présence de l'entrée dans le texte.

Dans cet article, nous considérons les fonctions suivantes:

- $Vote1\,(E_H, D) = \sum \frac{F(T,D)}{S(D)} . \frac{S(H)}{F(T,H)} . 10^{\frac{NT(E,D)}{NT(E)}}$

Cette fonction est notre fonction de "base". Elle accorde la même importance aux deux facteurs que sont l'importance du terme dans le document et l'importance du terme dans la HC. La présence de l'exponentielle pour le taux de présence de l'entrée dans le texte a pour but d'accentuer l'importance des entrées de HC suivant le nombre de termes de l'entrée présents dans le texte.

- $Vote2\,(E_H, D) = \sum \frac{F(T,D)}{S(D)} . 100 . 10^{\frac{NT(E,D)}{NT(E)}}$

Pour cette fonction (dite « sans SH ») le facteur mesurant l'importance d'un terme dans la HC est remplacé par une constante. L'objectif est d'analyser l'impact de ce facteur sur les résultats par rapport aux résultats obtenus avec la fonction de base.

- $Vote3\,(E_H, D) = \sum \frac{F(T,D)}{S(D)} . \frac{\sqrt{\log_{10} \frac{S(H)}{F(T,H)} + 1}}{e} . 10^{\frac{NT(E,D)}{NT(E)}}$

Cette fonction vise à relativiser l'influence du facteur mesurant l'importance d'un terme dans la HC par rapport au facteur mesurant l'importance du terme dans le document. En effet, lorsque les documents sont de grande taille, et particulièrement lorsque la taille de la HC est beaucoup plus petite, l'ordre de grandeur du facteur mesurant l'importance d'un terme dans la HC est beaucoup plus grand que celui du facteur mesurant l'importance du terme dans le document. Cette fonction limite également l'impact du taux de présence de l'entrée dans le texte.

- $Vote4\,(E_H, D, x) = \sum \frac{F(T,D)}{S(D)} . \frac{S(H)}{F(T,H)} . 10^{\frac{NT(E,D)}{NT(E)}}$ si au moins x% des termes de l'entrée apparaissent dans le document,

  = 0 sinon

Cette fonction accorde la prépondérance au taux de présence d'une entrée dans le texte. Les entrées constituées de termes dont peu apparaissent dans le document ont un score nul et ne font pas partie de l'ensemble des entrées associées au document. Pour les entrées ayant atteint le seuil fixé, la fonction de base est appliquée pour les ordonner. Le seuil est calculé par rapport à un pourcentage (couverture) de termes de l'entrée apparaissant dans le document. Une couverture de 100% indiquera que tous les termes de l'entrée doivent être présents dans le texte pour que l'entrée soit retenue. Pour une couverture de 50%, une entrée composée de 4 termes sera retenue si au moins 2 termes sont extraits du texte.

3.2 Cadre d'évaluation

La procédure d'association automatique a été comparée à une association réalisé manuellement. Pour cela, des évaluateurs ont été sollicités afin d'associer des documents Web (en langue anglaise et liés au domaine économique) selon trois HC : "Branch", qui regroupe les branches d'industries, "Country" qui regroupe les pays et "Theme" qui répertorie les indicateurs économiques d'intérêt. Ces HC définissent un domaine d'IRAIA. Les utilisateurs pouvaient associer jusqu'à 10 entrées d'une HC et indiquaient une force de représentation de 1 à 5.
Le jeu de test utilisé pour l'évaluation possédait les caractéristiques suivantes :
Nombre de documents : 40

HC "Branch" : 423 entrées	1570 termes	soit en moyenne un peu moins de 4 termes par entrée
HC "Country" : 23 entrées	25 termes	soit en moyenne environ 1 terme par entrée
HC "Theme" : 30 entrées	96 termes	soit en moyenne un peu plus de 3 termes par entrée

Les HC utilisées possèdent des caractéristiques différentes (nombres d'entrées différents, nombre de termes par entrée différents, nombre de niveaux différents). Elles sont représentatives des différents types de hiérarchies que l'on peut rencontrer.

3.3 Critères d'évaluation

Les critères d'évaluation que nous avons utilisés sont directement issus des critères utilisés pour évaluer les systèmes de recherche d'information : les taux de rappel et de précision. Dans notre étude, ces taux sont définis par:

$$Taux\ de\ rappel = \frac{Nombre\ d'entrées\ retrouvées\ et\ \ pertinentes}{Nombre\ d'entrées\ \ pertinentes}$$

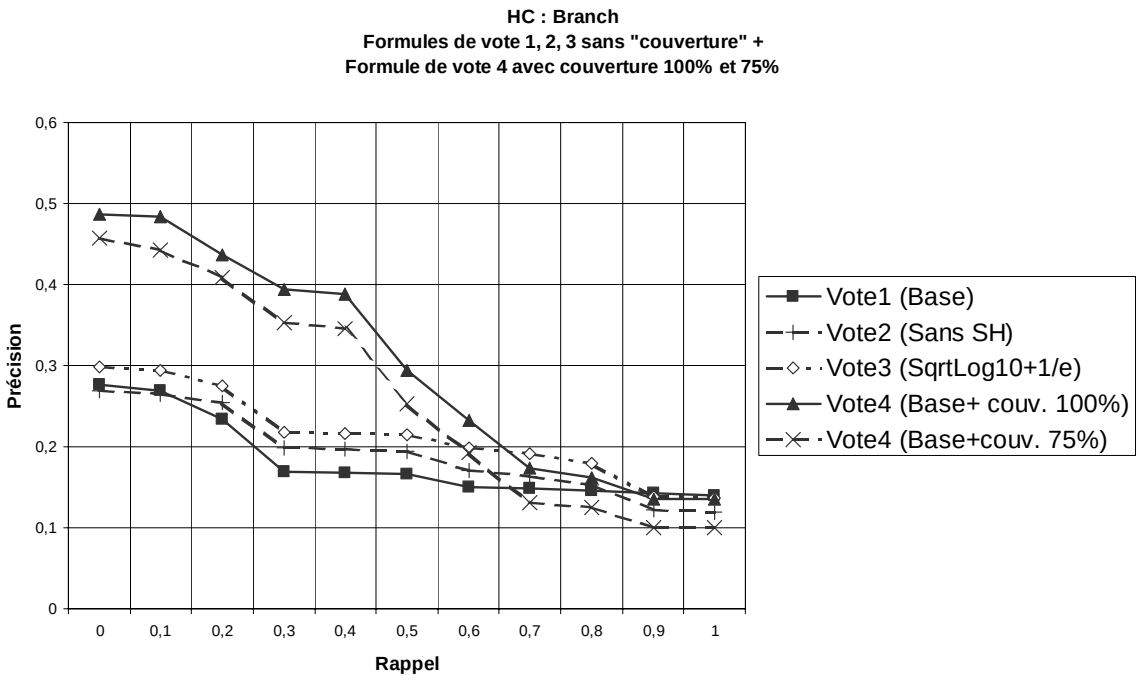
$$Taux\ de\ \ précision = \frac{Nombre\ d'entrées\ retrouvées\ et\ \ pertinentes}{Nombre\ d'entrées\ retrouvées}$$

Nous avons utilisé le programme trec-eval (cf. le site web trec.nist.gov) pour calculer les valeurs de rappel et de précision synthétisées dans les figures ci-après. Le taux de rappel et le taux de précision évoluent généralement de manière opposée. Le taux de précision est habituellement calculé pour différentes valeurs du taux de rappel (0, 0,1, 0,2, …, 1). La précision à N entrées est indiquée en plus de la R-précision. La R-précision mesure la précision après R entrées retrouvées où R est le nombre

total d'entrées pertinentes pour un document. La précision à N entrées correspond à la précision lorsque seules les N premières entrées retrouvées sont prises en compte. Les taux de précision à 5 entrées, 10 entrées, …, sont fournis par le programme trec_eval.
Il est à noter que la taille de l'ensemble d'éléments utilisé pour l'expérimentation, ici des documents, peut être considéré comme significatif. Dans le cadre du programme international TREC [10] par exemple, une cinquantaine éléments est généralement utilisée lors des évaluations.

3.4 Résultats et Discussion

Les figures suivantes synthétisent les mesures de rappel/précision pour chaque HC, ainsi que globalement pour toutes les HC confondues. Les résultats obtenus pour les fonctions évaluées sont indiqués. Une première figure détaille la précision à différentes valeurs du taux de rappel. Une seconde figure se focalise sur la précision à 5 et à 10 ainsi que sur la R-précision.



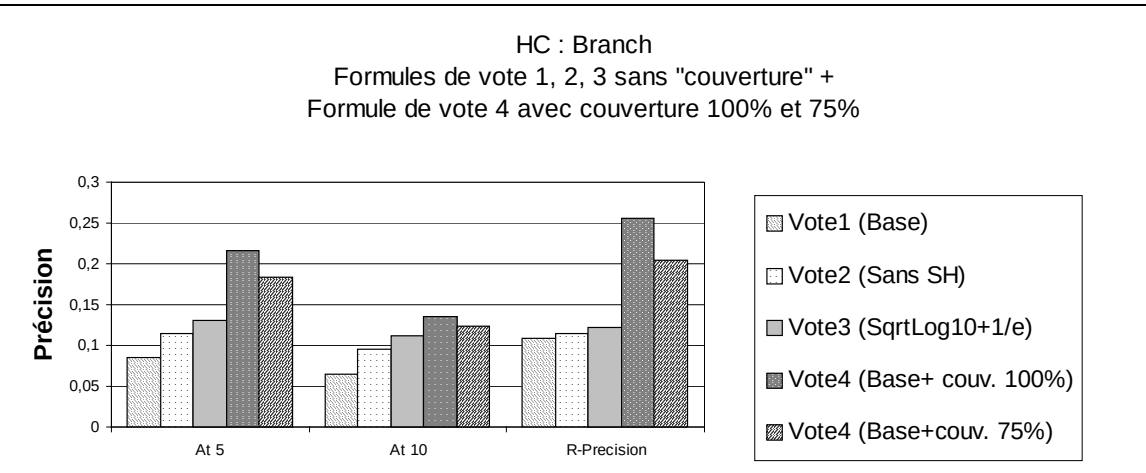


Figure 1.A.: Rappel/Précision considérant la HC "Branch" seulement.

Si l'on considère la hiérarchie "Branch", la prise en compte de la *couverture* influe de manière non négligeable sur les résultats.

Si l'on ne considère pas la couverture, les fonctions qui donnent les meilleurs résultats sont celles qui limitent la valeur du facteur S(H)/F(T,H) i.e. la taille de la HC divisée par la fréquence du terme dans la hiérarchie. Les taux de précision ne sont pas particulièrement satisfaisants (de 0,3 à 0,1), et ce quel que soit le taux de rappel. Néanmoins, ces résultats sont améliorés lorsque l'on considère la couverture. La précision est augmentée quel que soit le taux de rappel lorsqu'une couverture élevée est utilisée.
L'influence de la couverture sous-entend que, pour la hiérarchie "Branch", sans considérer de couverture, de nombreuses entrées pour lesquelles peu de termes apparaissent dans un document obtiennent des scores élevés. Ces entrées n'ont pourtant pas été jugées pertinentes de façon manuelle. L'introduction de la couverture élimine donc ce type d'entrée. Ce cas de figure peut apparaître de manière plus fréquente pour une hiérarchie telle que "Branch" qui possède un nombre moyen de termes par entrée plus élevé que les autres hiérarchies.

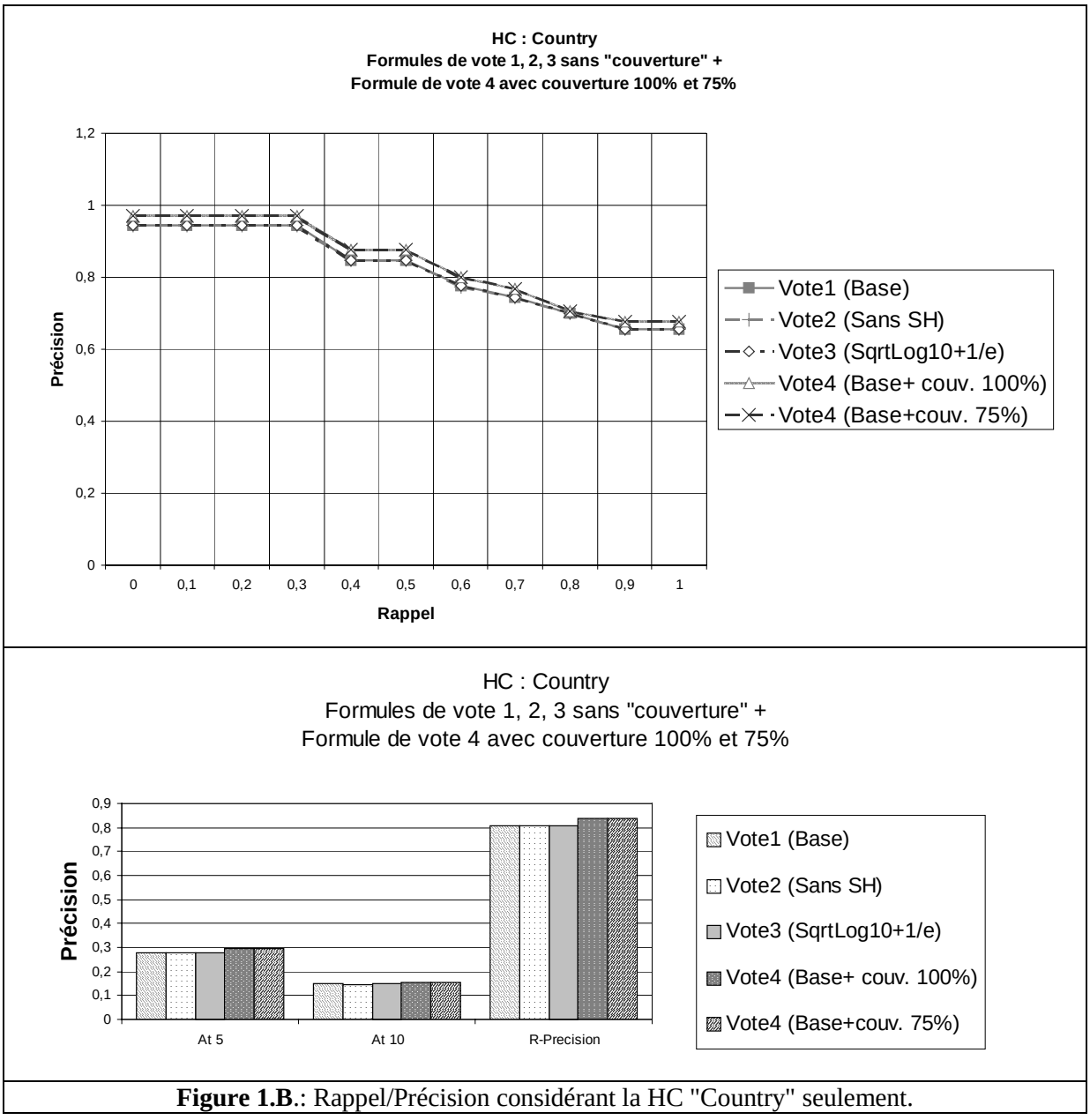


Figure 1.B.: Rappel/Précision considérant la HC "Country" seulement.

Si l'on considère la hiérarchie "Country", toutes les fonctions aboutissent pratiquement aux mêmes résultats. La précision est élevée quel que soit le taux de rappel (quasiment 1 pour les faibles taux de rappel et plus de 0,65 pour des taux de rappel élevés). La R-Précision est au-delà de 0,8. Ces très bons

résultats peuvent s'expliquer d'un part par le fait que la plupart des entrées de la HC "Country" ne sont composées que d'un seul terme, d'autre part, par le fait que les variations de termes utilisés pour indiquer un pays sont limités. La faible précision obtenue à 5 et 10 entrées peut être expliquée par le fait que généralement moins de 5 entrées ont été attribuées manuellement pour chaque document. En fait dans ce cas, la R-précision très élevée de 0,8 est un indicateur important. Elle signifie que si l'on considère les x premières entrées retrouvées sachant, x correspondant au nombre d'entrées effectivement pertinentes, 80% d'entre elles seront des entrées pertinentes.
Du point de vue de la couverture, la précision augmente encore légèrement, quel que soit le taux de rappel, lorsque l'on applique un pourcentage de couverture élevé.

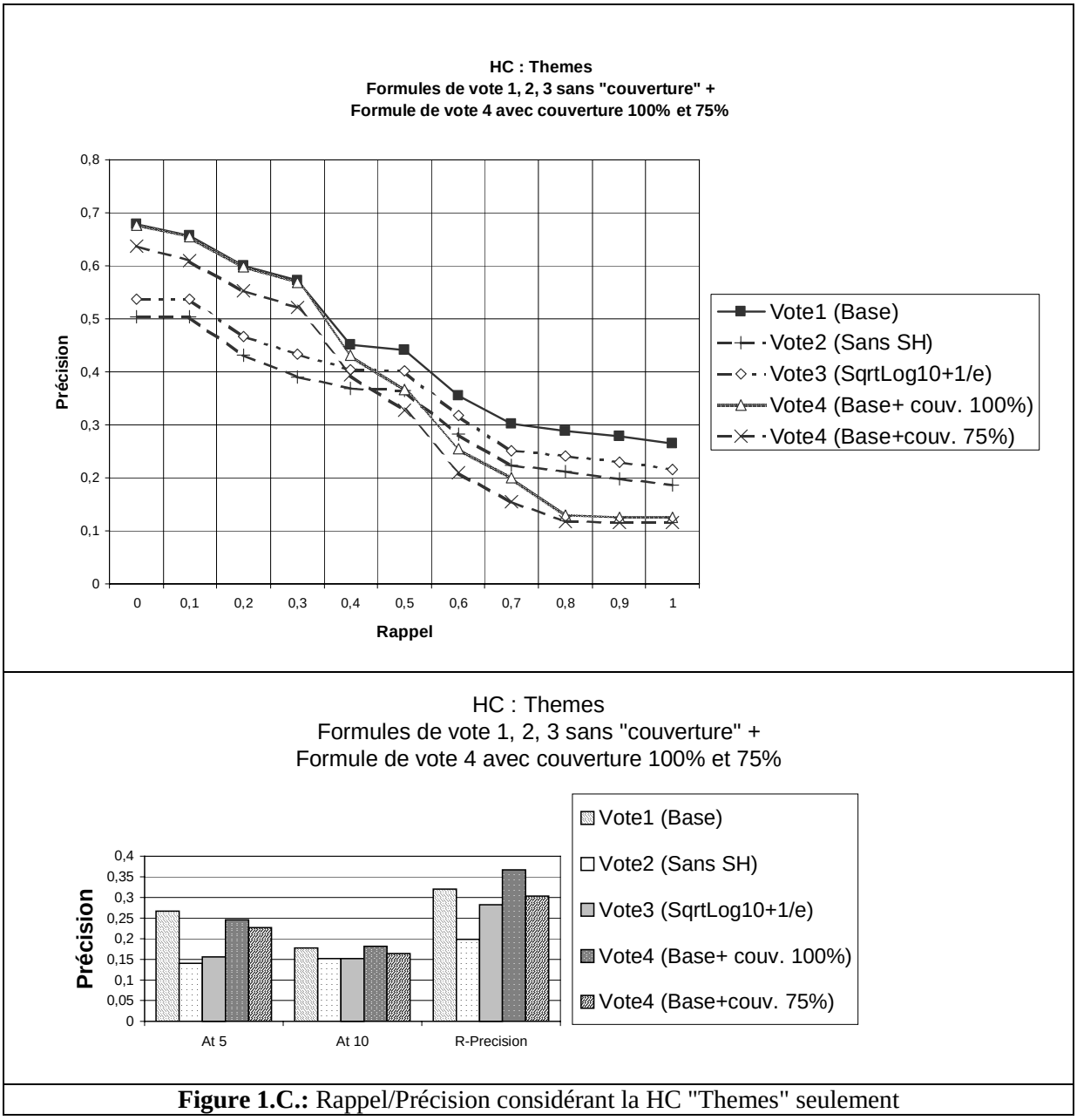


Figure 1.C.: Rappel/Précision considérant la HC "Themes" seulement

Si l'on considère la hiérarchie "Theme", la fonction Vote1 (de "base") fournit les meilleurs résultats. Au taux de rappel de 0,5, la précision est voisine de 0,45. La R-Précision est supérieure à 0,3.
A l'inverse de la hiérarchie "Branch", si l'on ne considère pas la couverture, les fonctions qui limitent la valeur du facteur S(H)/F(T,H) (i.e. taille de la HC divisé par la fréquence du terme dans la hiérarchie) donnent de moins bons résultats. Les différences de tailles entre les deux hiérarchies

(hiérarchie "Branch" 14 fois plus volumineuse que la hiérarchie "Theme") peuvent expliquer la différence de comportement de ce paramètre.
Contrairement à la HC "Branch" notamment, l'utilisation d'une couverture plus importante n'améliore pas les résultats. En effet, l'utilisation de la fonction de vote Vote4, que se soit avec une couverture de 75% ou de 100%, donne des taux de précision inférieurs à ceux produits par la fonction de vote Vote1, plus particulièrement pour des taux de rappel supérieur à 0,4. Une couverture faible est donc meilleure, puisque l'on peut considérer que la fonction de vote Vote1 est équivalente à la fonction Vote4 avec une couverture minimum c'est-à-dire de 0%. De plus, une expérimentation non présentée dans les figures de la fonction Vote4 avec une couverture de 10% conforte cette conclusion.

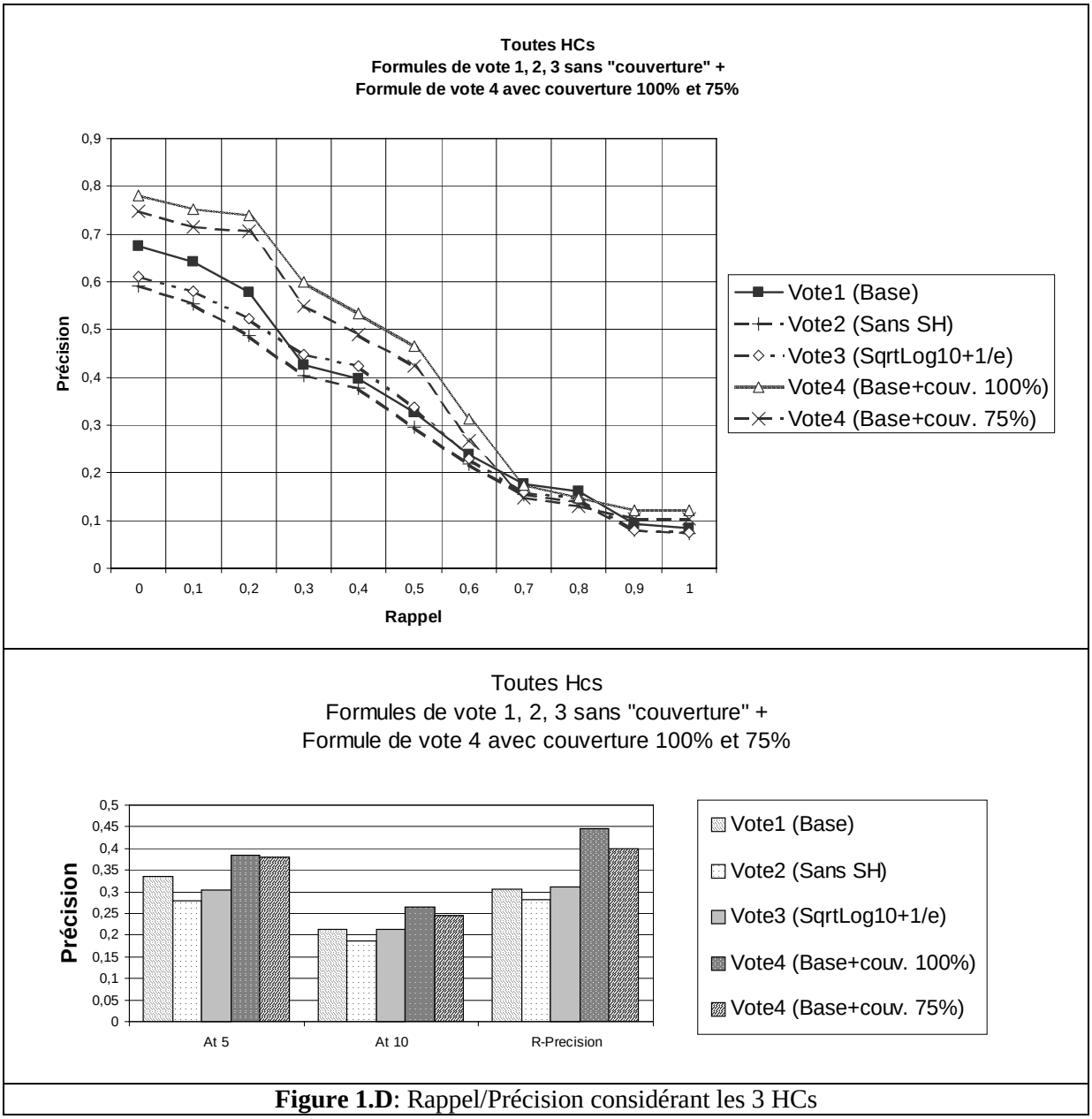


Figure 1.D: Rappel/Précision considérant les 3 HCs

En considérant globalement les 3 hiérarchies indifféremment, la fonction Vote4 (de "base") fournit les meilleurs résultats. De plus, il semble que plus la couverture est grande, meilleurs sont les résultats.

4 Conclusion

Nous avons proposé un mécanisme d'association de textes à des hiérarchies de concepts. Ce principe consiste à identifier les entrées des hiérarchies de concepts les plus représentatives du contenu d'un texte. L'importance d'une entrée par rapport à un texte est calculée par une méthode de vote. Le principe d'association entre textes et hiérarchies a été mis en œuvre dans le cadre du projet européen IRAIA. IRAIA est un système permettant une recherche contextuelle d'information dans de larges espaces de données. L'association automatique des données aux hiérarchies permet de créer des espaces d'informations bien identifiés pour lesquels les hiérarchies de concepts servent à la formulation du besoin d'information. L'utilisateur évolue dans un contexte sémantique de recherche dont il ne peut s'éloigner, palliant le problème d'ambiguïté rencontré par les moteurs de recherche traditionnels. Le mécanisme proposé a été évalué sur une collection de test, avec différentes fonctions de vote. Les résultats ont montré que :

- concernant la HC de localisation géographique la précision reste au dessus de 80%, les différentes fonctions permettent d'obtenir les mêmes types de résultats,
- concernant les HC de thèmes et de branches, certaines fonctions permettent d'obtenir de meilleurs résultats. Ces fonctions diffèrent selon la hiérarchie.

Nous allons poursuivre l'évaluation en particulier en nous basant sur des collections de grandes tailles utilisées en classification automatique. (collections Ohsumed -ftp://medir.ohsu.edu/pub/ohsumed/ et Reuters - ftp://ciir-ftp.cs.umass.edu/pub/).

5 References

[1] J. Allen, Making a semantic Web, http://www.netcrucible.com/semantic.html

[2] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American, 2001, http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html.

[3] K. Englmeier, J. Mothe, Thustworthy personal assistance: a design objective agents, Association for information systems, 7th Americas Conference on Information Systems, (Cédérom), Boston, Août 2001.

[4] M.A. Hearst, Next generation Web search: setting our sites, IEEE Data Engineering Bulletin, Vol.23, N.3, pp 38, 2000.

[5] IRAIA, Projet soutenu par la commission Européenne via le 5ième programme cadre, Getting Orientation in Complex Information Spaces as an Emergent Behavior of Autonomous Information Agents, IST-1999-10602.

[6] A. James, A.M. Day, Web clustering: a new approach to space partitioning, *WSCG'99 Conference Proceedings* Volume 1, pg 165-172, 1999.

[7] Y. Li, Toward a qualitative search engine, IEEE Internet Computing, pp 24-29, Vol.2, N.4, 1998.

[8] B. Pauer, P. Holger, Statfinder, Document Package Statfinder, Vers. 1.8, mai 2000.

[9] Ali Asghar Shiri, Crawford Revie, Thesauri on the Web: current developments and trends, Online information review, Vol 24, N.4, pp273-279, 2000.

[10] TREC, Text Retrieval Conference, http://trec.nist.gov.

6 Remerciements

Les recherches présentées dans ce papier s'inscrivent dans le cadre du projet Européen IRAIA soutenu par la Commission Européenne dans le 5^{ième} programme cadre (IST-1999-10602). Cependant, les idées exprimées dans ce papier nous sont personnelles et ne correspondent pas nécessairement à celles du consortium.