Evaluation de la pertinence dans un système de recommandation sémantique de nouvelles économiques

David Werner, Christophe Cruz, Aurélie Bertaux

LE2I UMR 6306
Faculté des sciences Mirande
BP 47870, 21078 Dijon Cedex, France
david.werner,christophe.cruz,aurelie.bertaux@u-bourgogne.fr

Résumé. De nos jours dans les secteurs commerciaux et financiers la veille est cruciale et complexe, car la charge d'informations est importante. Pour répondre à cette problématique, nous proposons un système novateur de recommandation d'articles basé sur une modélisation ontologique des connaissances. Nous présentons également une nouvelle méthode d'évaluation de la pertinence utilisant le modèle vectoriel intrinsèquement efficace et adapté afin de pallier la confusion native de ces modèles entre les notions de *similarité* et de *pertinence*.

1 Introduction

La veille économique s'inscrit aujourd'hui pleinement dans la stratégie de développement des entreprises. Or, la quantité d'informations à leur disposition est considérable, rendant l'analyse complexe. L'entreprise partenaire de ces travaux publie quotidiennement des articles synthétisant des informations économiques émanant de différentes sources ou fruit d'une démarche d'investigation. Afin de les adresser au mieux aux lecteurs concernés, nous développons un outil efficace de recommandation de ces articles d'actualités économiques régionales, reposant sur son adéquation avec les besoins des utilisateurs. Pour personnaliser la recommandation, une enquête a été menée auprès des clients avec l'appui des experts du domaines. Cela nous a permis d'identifier trois critères principaux sur lesquels nous nous appuyons : les *Thèmes* (principaux événements économiques), les *Secteurs* d'activité et la *Localisation* des informations. Les besoins des utilisateurs et le contenu informationnel des articles sont représentés par une description sémantique des connaissances de ces trois critères au sein d'une ontologie. Dans cet article, nous nous intéressons à la distinction entre *pertinence* et *similarité*, qui sont souvent amalgamés. Nous proposons donc une nouvelle mesure, *Relevancy Measure*, nous permettant de définir la pertinence d'un article pour un profil donné en nous appuyant sur leurs descriptions ontologiques. Nous utilisons un système de recommandation basé sur le contenu avec une approche vectorielle. Cet article est organisé de la façon suivante : nous commençons par présenter la génération des vecteurs, puis nous introduisons les notions de similarité et de pertinence et définissons la mesure *Relevancy Measure*. Enfin, la section 4 propose une évaluation de nos algorithmes avant la présentation de nos conclusions.

2 Vectorisation

La description de chaque article et profil est contenue dans une base de connaissances ontologique. Afin d'utiliser le modèle vectoriel, il est nécessaire de transformer ces descriptions en vecteurs. Cette modélisation est bien moins expressive qu'une ontologie car les dimensions étant orthogonales dans le modèle vectoriel, tous les éléments de chaque vecteur sont considérés comme indépendants (Voorhees, 1994).

Génération des vecteurs : La mise en relation d'un article avec les critères qui lui sont associés est réalisée de façon semi-automatique via la plate-forme GATE (Cunningham, 2002). Les résultats de l'analyse du texte par cette plateforme sont vérifiés, corrigés et validés par les rédacteurs de chacun des articles puis remontés dans l'ontologie. La création des profils est réalisée manuellement lors de l'inscription des clients. Les vecteurs décrivant articles et profils contiennent les critères qui leur ont été associés dans la base de connaissances ontologique.

Expansion des vecteurs : D'après ce qui a été dit précédemment, les vecteurs de description des articles et profils ne contiennent que les instances de critères avec lesquels ils sont directement en relation dans la base de connaissances. Nous nous basons sur les apports de (IJntema et al., 2010), dont nous avons adapté la méthode dite d'*expansion de vecteurs* afin de conserver les connaissances de l'ontologie dans les vecteurs. Les instances de chaque critère sont organisées de façon hiérarchique dans la base de connaissances. Pour chaque instance ajoutée aux vecteurs, les instances parents y sont elles aussi ajoutées.

3 Similarité versus Pertinence

Similarité : $Similarite(x, y) : I \times I \rightarrow [0, 1]$ est une fonction qui permet d'évaluer le degré de similarité entre deux objets x et y. Dans notre cas, x est un article et y un profil. Cette fonction doit satisfaire les trois propriétés de *positivité*, *réflexivité* et *symétrie*. L'évaluation de la similarité dans un espace vectoriel peut être réalisée par différentes mesures, similarité Cosinus, similarité Jaccard, ou distance euclidienne pour ne citer que les plus utilisées. Dans cet article, nous illustrons notre propos avec la similarité cosinus. La similarité Cosinus entre deux vecteurs $\overrightarrow{a}$ et $\overrightarrow{p}$ repose sur la mesure du cosinus de l'ange Θ entre les deux vecteurs.

Pertinence et Relevancy : $Pertinence(x, y) : I \times I \rightarrow [0, 1]$ est une fonction qui permet de mesurer le degré de pertinence d'un article x vis-à-vis d'un profil y. Cette mesure de pertinence doit aussi respecter les propriétés de positivité et réflexivité. La pertinence est une notion largement utilisée dans le domaine de la recherche d'informations. Dans notre cas, la pertinence n'est pas binaire. Un article peut plus ou moins correspondre au besoin d'informations d'un utilisateur, c'est pourquoi nous utilisons le modèle vectoriel pour l'estimer. Contrairement aux approches classiques confondant les notions de similarité et de pertinence (Salton, 1971), nous les distinguons. Par exemple, si un profil montre un intérêt pour la Côte d'Or, il est préférable de lui recommander un article plus précis, qui traite de Dijon, qu'un article plus général qui traite de la Bourgogne. Il faut donc conserver une pertinence forte dans le sens de la spécialisation et la baisser fortement dans le sens de la généralisation du critère. Afin de résoudre

ce problème, nous utilisons un vecteur intermédiaire. Le sous-vecteur $\overrightarrow{s_c}$ est composé des instances communes entre le vecteur de l'article $\overrightarrow{a_c}$ et celui du profil $\overrightarrow{p_c}$. Ainsi, nous définissions la pertinence pour un critère donné *c* de la façon suivante :

$$Pertinence_c(\overrightarrow{a_c}, \overrightarrow{p_c}) = \frac{\omega'_{1,c} \times Similarite_c(\overrightarrow{a_c}, \overrightarrow{s_c}) + \omega'_{2,c} \times Similarite_c(\overrightarrow{p_c}, \overrightarrow{s_c})}{\omega'_{1,c} + \omega'_{2,c}}$$

Avec S_c le sous-ensemble commun d'éléments de l'ensemble d'instances en relation à la fois avec le profil $I'_{p,c}$ et l'article $I'_{a,c}$; $S_c = I'_{p,c} \cap I'_{a,c}$. $\forall i_{x,c} \in S_c$, le vecteur $\overrightarrow{s_c}$ est composé des éléments de l'ensemble S_c ; $\overrightarrow{s_c} =< i_{1,c}, i_{2,c}, ..., i_{t,c} >$. Avec cette méthode, il est possible de pondérer l'importance de la différence de précision entre profils et articles. Dans notre cas, nous utilisons $\omega'_{1,c} = 1$ et $\omega'_{2,c} = 4$, car nous considérons que la perte de précision du profil par rapport à l'article ne doit pas influencer plus de 20% du résultat. De plus, la perte de précision de l'article par rapport au profil doit influencer fortement le résultat, ici 80%. La pertinence globale $Relevancy(\overrightarrow{a}, \overrightarrow{p})$ est la somme des mesures de pertinence pour chacun des critères, éventuellement pondérées. Cette mesure est utilisée dans notre prototype pour trier les résultats (articles) proposés à l'utilisateur en fonction de son profil :

$$Relevancy(\overrightarrow{a}, \overrightarrow{p}) = \frac{\sum \omega_c * Pertinence_c(\overrightarrow{a_c}, \overrightarrow{p_c})}{\sum \omega_c}$$

4 Expérimentations

Nous avons comparé de deux façons différentes (évaluation binaire et d'ordre) les résultats de la recommandation d'articles via les méthodes, similarité cosinus (*C*), similarité cosinus avec vecteur étendus (*B*) et *Relevancy Measure* avec vecteur étendus (*A*), qui permet de gérer la différence de précision entre les profils et les articles. Pour nos évaluations, un ensemble de 10 profils et 70 articles a été choisi. Pour l'évaluation binaire, une sélection manuelle des articles pertinents a été réalisée pour chaque profil par des experts. Pour l'évaluation de l'ordre, un classement manuel des articles pertinents a été réalisé pour chaque profil par des experts. Dans les deux cas, le travail des experts est considéré comme la recommandation idéale, avec laquelle sont comparés les résultats des différents algorithmes.

Evaluation Binaire : Pour évaluer la recommandation binaire nous utilisons les mesures classiques en recherche d'informations, *précision*, *rappel* et *F1-mesure*, (Lewis et Gale, 1994) qui produisent des scores allant de 0 à 1. Tous les articles dont la corrélation avec le profil est supérieure à un seuil de 0,5 sont conservés. Les résultats de l'évaluation de la recommandation

Algorithmes	Précision	Rappel	F1-mesure	Tau de Kendall	Rho de Spearman
A	**0.856**	**0.971**	**0.910**	**0.837**	**0.899**
B	0.916	0.453	0.607	0.830	0.894
C	0.883	0.181	0.301	0.713	0.694

TAB. 1 – *Résultat des mesures d'évaluation binaires et de rang pour chaque algorithme.*

binaire présentés dans la table 1 confirment les résultats de (Voorhees, 1994) quant à l'intérêt

de l'*expansion de vecteurs* et montrent que notre méthode *Relevancy Measure*, distinguant la pertinence de la similarité directe, fournit le meilleur résultat.

Evaluation de l'ordre : Pour évaluer l'ordre de recommandation des articles, nous utilisons les deux mesures de corrélation linéaire de rang les plus populaires : le rho de Spearman et le tau de Kendall. Ces deux mesures produisent des scores allant de -1 à 1. 0 étant l'absence de similitude, 1 la similitude complète et -1 l'inverse. Les résultats de l'évaluation sont présentés dans la table 1. L'intérêt de l'*expansion de vecteurs* est encore une fois confirmé. De plus, les résultats montrent que l'utilisation de la *Relevancy Measure* améliore davantage les performances du système.

5 Conclusion

Dans cet article, nous avons distingué la *Similarité* de la Pertinence. Nous avons défini une mesure de pertinence (*Relevancy Measure*) entre un article et un profil sémantiquement décrits dans une ontologie et adaptée au modèle vectoriel. Cette mesure permet de prendre en compte le degré de précision dans la description du besoin, lors de l'évaluation de la pertinence. Nous avons démontré que cette mesure donne de bons résultats pour la recommandation d'articles. Nous projetons une évaluation sur un jeu de données plus vaste, prenant également en compte le comportement des utilisateurs.

Références

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities 36*(2), 223–254.

IJntema, W., F. Goossen, F. Frasincar, et F. Hogenboom (2010). Ontology-based news recommendation. pp. 1. ACM Press.

Lewis, D. D. et W. A. Gale (1994). A sequential algorithm for training text classifiers. In *SIGIR '94*, pp. 3–12. Springer London.

Salton, G. (1971). The SMART retrieval system - experiments in automatic document processing.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *SIGIR '94*, pp. 61–69. Springer London.

Summary

Today in the commercial and financial sectors, staying informed about economic news is crucial and complex because of the huge amount of information. To address this problem, we propose an innovative article recommender system based on a knowledge ontological model. We present also a novel method to evaluate the relevancy based on vector space model that we have perfected to overcome the mix up existing in models between the concepts of *similarity* and *relevancy*.