Classification de documents en réseaux petits mondes en vue d'apprentissage

Khazri Mohamed*, Tmar Mohamed**, Mohand Boughanem***, Abid Mohamed*
*Ecole Nationale d'Ingénieurs de Sfax, Route de Soukra, 3038, Sfax
Mohamed.khazri@yahoo.fr
mohamed.Abid@enis.rnu.tn
**Institut Supérieur d'Informatique et du Multimédia de Sfax, 3018, Sfax
mohamed.tmar@isimsf.rnu.tn
***Institut de Recherche en Informatique de Toulouse, route de Narbonne, 31000, Toulouse, France
bougha@irit.fr

1 Introduction

Les systèmes de recherche d'information préconisent une fonctionnalité très intéressante voire indispensable lors de tout processus de recherche : il s'agit de la reformulation automatique de la requête. Cette fonctionnalité permet de rétablir les choix de l'utilisateur dans la perspective de retrouver plus de documents qui répondent à son besoin en information. Il est à noter à ce niveau que le besoin en information de l'utilisateur est très vague : l'utilisateur ne sait en général pas ce qu'il cherche. Par ailleurs, il peut tolérer un résultat initial imprécis sous réserve de l'améliorer par feedback Rocchio (1971).

Faire recours à de nouvelles méthodes d'apprentissage est alors devenu une nécessité. Plusieurs modèles qui ont été auparavant délaissés, tels que la classification, sont repris en vu d'améliorer l'apprentissage en recherche d'information. Nous proposons dans ce papier une méthode d'apprentissage en faisant appel aux réseaux petits mondes (*small worlds en anglais*, Watts (1999)).

2 Notre Approche

Les propriétés des réseaux petits mondes paraissent intéressantes dans les problèmes de classification. D'autant plus que ces propriétés sont valuées. Comme application à la recherche d'information, nous présumons qu'un ensemble de documents peut constituer des réseaux petits mondes pour moins qu'ils parlent du même sujet, et qu'une idée peut être transmise d'un document à un autre document si les auteurs partagent le même intérêt.

Nos objectifs pour l'intégration des small worlds en recherche d'information ont deux effets : un effet de construction des small worlds par le bais de la classification; et un effet d'estimation de pertinence sur d'autres documents.

En partant de l'hypothèse suivante : «*une classe est raisonnable si elle admet certaines propriétés : celles des small worlds*». Le premier effet va simplement faire une construction de small worlds de documents homogènes (pertinents ou non pertinents). Pour ce faire, nous proposons trois stratégies : une stratégie de construction de graphes de documents (1), une stratégie de propagation des liens (2), et une stratégie de construction des classes des documents (3). Pour la stratégie (3) nous utiliserons une méthode de classification hiérarchique, et l'identification du nombre de classes dépend de la qualité de classification et de la nature de

classes construites. A chaque itération nous calculons une valeur d'inertie intra-classe qui permet de quantifier l'homogénéité de la classification. Pour des classes réellement construites les coefficients de clustérisation et les distances moyennes montrent que les classes construites admettent les propriétés des small worlds.

En partant de l'hypothèse suivante : «*une classe est un small worlds, et qu'une classe homogène (constituée de documents pertinents ou non pertinents) peut être utilisée comme moyen efficace pour bien constituer l'estimation des scores d'autres documents*», le deuxième effet consiste à estimer la pertinence pour d'autres documents. Pour traduire la pertinence pour un document il suffit d'identifier la classe à laquelle il appartient et de juger de sa pertinence en fonction de la nature de la classe. Ce document est jugé pertinent si la classe résultat contient plus de documents pertinents que de documents non pertinents et est jugé non pertinent si non,

3 Conclusion

Nous avons présenté dans cet article une approche statistique de classification des documents. L'approche consiste à définir un nouveau concept d'apprentissage. L'apprentissage consiste à construire des classes qui préservent les propriétés des réseaux petits mondes. Nous admettons que les classes préservant ces propriétés sont des estimateurs de pertinences d'autres documents. L'approche que nous avons proposée consiste à considérer tous les critères pouvant intervenir dans le jugement de l'utilisateur et de leur affecter les meilleurs poids pour que la pertinence utilisateur soit proche de la pertinence système.

Les poids des critères considérés sont ajustés par apprentissage. Chaque poids traduit l'intérêt porté par l'utilisateur à celui-ci. Les poids relatifs aux termes peuvent servir de moyen de construction de requête. Nous envisageons de tester l'approche sur une base réelle de documents afin de mesurer l'apport des réseaux petits mondes à la recherche d'information. Nous envisageons également de tester la reformulation de la requête en se basant sur les poids des critères. Avec l'effet petit monde, nous envisageons d'autres méthodes telles que les méthodes d'ordonnancement (***ranking effect***).

Références

Watts, D.J. (1999). Small Worlds. *Princeton university press*. Princeton.

Rocchio, J. (1971). Relevance feedback in information retrieval. *SMART retrieval-system: experiments in automatic document processing*, 313-323.

Summary

This paper presents a statistical approach to classify a corpus of documents. The corpus is represented by a graph where nodes are represented by the documents and links are defined by some criteria. The classification aims to build homogenous small worlds (containing as much as possible only relevant documents or non relevant documents).These classes are used to estimate the scores of other documents.