Recherche d'information avec des modèles de langue

Jian-Yun Nie

DIRO, Université de Montréal
CP. 6128, succursale Centre-ville, Montréal, Québec,
H3C 3J7 Canada

Résumé : Des modèles de langue statistiques ont été développés dans la linguistique informatique. Ces modèles tentent de capter les régularités d'une langue en observant les occurrences des mots ou des suites de mots dans un corpus d'entraînement. Une fois un modèle entraîné, nous pouvons déterminer la probabilité d'une séquence quelconque de mots dans cette langue, selon le modèle.

Récemment, les approches basées sur des modèles de langues ont été utilisées avec succès en recherche d'information (RI). Une différence notable avec les approches traditionnelles à la RI est que les approches de modèle de langue ne modélisent généralement pas explicitement la notion de pertinence. En général, ces approches tentent de construire un modèle de langue M_D pour chaque document. La correspondance d'un document à une requête Q sera déterminée par la probabilité que la requête puisse être générée par le modèle du document, i.e. $P(Q|M_D)$.

Dans cette présentation, nous passons en revue les principales approches développées pour la modélisation de langue en linguistique informatique. Un aspect important dans cette modélisation est le lissage, qui permet d'attribuer une probabilité non nulle aux mots ou aux séquences de mots non rencontrés dans le corpus d'entraînement. Le lissage joue un rôle important dans l'application des modèles de langue en RI, car beaucoup de mots seront absent du document D, et du modèle M_D. Ainsi, le modèle du document M_D doit être lissé. Ce lissage est souvent effectué en combinant le modèle du document avec un modèle du corpus qui assure une meilleure couverture.

Ensuite, nous présentons les différentes approches proposées dans la littérature, qui utilisent des modèles de langues pour la RI. Les résultats expérimentaux seront également décrits.

Finalement, nous comparons les approches basées sur des modèles de langue avec quelques approches plus traditionnelles, tels que les modèles probabilistes, le modèle vectoriel et les modèles logiques. Nous conclurons en mentionnant quelques problèmes restant dans les modèles de langues, et des avenues possibles dans leur développement futur.

XML retrieval: from model to evaluation.

Mounia Lalmas

Department of Computer Science
Queen Mary University of London,
London E1 4NS, United Kingdom
mounia@dcs.qmul.ac.uk

The widespread use of the extensible Markup Language (XML), especially the increasing use of XML in scientific data repositories, Digital Libraries and on the Web, has brought about an explosion in the Development of XML tools, including systems to store and access XML content. The aim of such retrieval systems is to exploit the logical structure of documents, which is explicitly represented by the XML markup, to retrieve document components instead of whole documents in response to a user's query. Implementing this more focused retrieval paradigm means that an XML retrieval system needs not only to find relevant information in the XML documents, but also determine the appropriate level of component granularity to return to the user. In addition, the relevance of a retrieved component is dependent on meeting both content and structural conditions.

This talk will start with an overview of the issues involved with the effective content-oriented retrieval of XML documents. It will then present groups of XML retrieval systems that have been developed, and in particular how they deal with the aforementioned issues. The talk will continue with a description of INEX, the Initiative for the Evaluation of XML Retrieval, which provides an opportunity for participants to evaluate their XML retrieval methods using uniform scoring procedures.

The talk will end with a list of open issues regarding the development and evaluation of XML retrieval systems.