وزارة التعليم العالي والبحث العلمي

BADJI MOKHTAR UNIVERSITY-ANNABA
-UNIVERSITE BADJI MOKHTAR-ANNABA-

جامعة باجي مختار
- عنابة -

Faculté : Sciences de l'Ingénieur **Année 2008**
Département : Informatique

BIBLIOTHEQUE DU CERIST

MÉMOIRE

Présentation en vue de l'obtention du diplôme de magister

Recherche et filtrage d'information basés sur le Text Mining sous la technologie GRID

Option : Texte Parole et Image

Présenté par : BOUALI Ali

DEVANT LE JURY

Pr BADACHE Nadjib, Professeur à l'USTHB Alger, Président
Pr LASKRI Mohamed Tayeb, Professeur à l'Université de Annaba, Rapporteur
Dr FARAH Nadir, Maitre de Conférence à l'Université de Annaba, Examinateur
Dr BOUHADADA Tahar, Maitre de Conférence à l'Université de Annaba, Examinateur

Résumé

La quantité d'information textuelle augmente de façon exponentielle aussi bien comme archives que documents de travail dans les organisations académiques, dans les administrations et dans les entreprises.

Il est très difficile aux engins de recherche de retrouver l'information adéquate dans cet important volume de données textuelle, reflétant les intérêts des utilisateurs qui changent avec le temps. Nous avons donc besoin de techniques d'apprentissage permettant de reconnaître les intérêts des utilisateurs en donnant uniquement une simple requête, et de filtrer l'information disponible suivant ces intérêts, et donc de mettre en place un système de recherche et de filtrage d'informations.

Dans ce travail, on propose une architecture du système de recherche et filtrage d'information basé sur le Text Mining sous la technologie GRID, capable de constituer et consulter une base de données d'une part, et de filtrer l'information disponible suivant les besoins des utilisateurs d'une autre part.

Le mécanisme de filtrage est à la base du Text Mining, pour le quel on a appliqué les techniques du GRID pour améliorer les performances de calcule, il réalise la modélisation des intérêts des utilisateurs et, le filtrage d'information; en effet, ces informations sont d'abord proposés par la rétroaction de l'utilisateur, et des mots sont ensuite sélectionnés pour créer un profil. En utilisant ce profil, l'information entrante est filtrée et ainsi, plus d'informations pertinentes sont présentée à l'utilisateur associé.

L'architecture du système de recherche et filtrage d'information basé sur le Text Mining sous la technologie GRID (RFITM-GRID) que nous avons proposé se compose de plusieurs modules :

BIBLIOTHEQUE DU CERIST

Un module d'indexation, qui, en balayant un ou plusieurs fichiers, il construit un index avec les mots trouvés dans les documents.

Un module moteur de recherche, qui permet de réaliser une recherche d'information en consultant les fichiers de la base de donnée créer par le module d'indexation.

Un module de filtrage, qui, à base de la rétroaction de pertinence et du Text Mining, en exploitant les hautes performances du GRID, crée des profils, qu'il exploite par la suite pour filtrer l'information entrante.

BIBLIOTHEQUE DU CERIST

Mots clés

Recherche d'information, Grille de calcule (GRID), Indexation, Rétroaction de pertinence, filtrage d'information, Profil utilisateur, découverte des connaissances dans les bases de données (DCBD), Text Mining.

Abstarct

The amount of textual information increases exponentially as well as archive papers in academic organizations, government and enterprise.

It is very difficult for search engines to find adequate information in this important volume of text data, reflecting the interests of users change with time. We need learning techniques to recognize users' interests by giving only a simple query and filter information available following these interests, and thus to establish a system of search and filtering information .

In this work, it proposes a system architecture search and filtering information based on Text Mining in the GRID, able to constitute and consult a database on the one hand and filters information available following user needs on another hand.

The filtering mechanism is the basis of Text Mining, for which we applied the GRID technology to improve performance calculates, it models the user's interests and achieves the filtering information, because this information are first proposed by feedback from the user, and key terms are then selected to create a profile. By using this profile, the incoming information is filtered and thus more relevant information is displayed to the user partner.

The system architecture search and filtering information based on Text Mining in the GRID technology (RFITM-GRID) that we have proposed consists of several modules:

A module indexing, this, by browsing one or several files, built an index with words found in the documents.

A search engine module, which allows a search for information by consulting the files of the database created by the indexing module.

BIBLIOTHEQUE DU CERIST

A filtering module, which, based on the relevance feedback and Text Mining, exploiting the high performance GRID, create profiles witch will be used to filter the new information.

Key words

Information retrieval, GRID, Indexing, Relevance feedback, information filtering, User profile, Knowledge discovering in data base (KDD), Text Mining.

BIBLIOTHEQUE DU CERIST

ملخـــص

عدد معلومات نصية يزيد أضعافا مضاعفة فضلا عن أوراق في أرشيف الأكاديمية والمنظمات والمؤسسات الحكومية.

ومن الصعب جدا لمحركات البحث للعثور على معلومات كافية في هذا النص المهم من حجم البيانات، والتي تعكس مصالح المستخدمين المتغيرة مع الوقت. ونحن بحاجة إلى تعلم تقنيات الاعتراف مصالح المستخدمين بإعطاء مجرد استفسار والمعلومات المتاحة فلتر بعد هذه المصالح، وبالتالي لإنشاء نظام للبحث وترشيح المعلومات.

في هذا العمل ،نقترح هيكل نظام البحث وترشيح المعلومات على أساس نص التعدين في غريد ، وتكون قادرة على خلق والتشاور قاعدة بيانات من جهة و ترشيح المعلومات المتاحة وفقا لاحتياجات المستخدمين من جهة أخرى.

فإن آلية الترشيح هو أساس نص التعدين ، عليه نطبق تكنولوجيا غريد لتحسين أداء الحساب ، يؤدي نمذجة المستخدمين ، وترشيح المعلومات ، لأن هذه المعلومات هي الأولى التي اقترحها التغذية المرتدة من المستعملين ، واختيار الكلمات بعد ذلك لخلق ملف.
باستخدام هذا الملف، فإن المعلومات الواردة ترشح وبالتالي المزيد من المعلومات ذات الصلة تعرض إلى المستخدم المرتبطة به.

بنية النظام والبحث وترشيح المعلومات على أساس نص التعدين في غريد (RFITM-GRID) الذي اقترحنها تتكون من عدة وحدات :

وحدة الفهرسة، التي، عن طريق المسح واحدة أو أكثر من ملفات، انه يبني فهرس بالكلمات التي تم العثور عليها في وثائق.

BIBLIOTHEQUE DU CERIST

وحدة محرك البحث ، الذي يتيح البحث عن المعلومات من خلال التشاور مع ملفات قاعدة البيانات التي أنشأتها وحدة الفهرسة.

وحدة الترشيح ، التي ، على أساس التغذية المرتدة و نص التعدين ، واستغلال الأداء العالي للغريد ، وخلق ملفات ، تعمل بعد ذلك لترشيح المعلومات الواردة.

BIBLIOTHEQUE DU CERIST

الكلمات الدالة

استرجاع المعلومات ، الحوسبة الشبكية ، الفهرسة ، وأهمية التغذية المرتدة ، وترشيح المعلومات ،ملف المستخدم ، واكتشاف المعرفة في قواعد البيانات ، نص التعدين.

Table des Matières

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST