REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

N° d'ordre : .....

UNIVERSITE DE M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

Département d'Informatique

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Systèmes d'Informations Avancés / Réseaux

Par: DJAIDJAA Imen

SUJET

Un portail informationnel basé sur le parsing et l'extraction documentaire

Soutenu publiquement le : / /2014 devant le jury composé de :

.....	Université de M'sila	Président
Brahimi Mahmoud	Université de M'sila	Rapporteur
.....	Université de M'sila	Examinateur
.....	Université de M'sila	Examinateur

Promotion : 2013 /2014

Sommaire

Introduction générale :

Ce travail s'inscrit dans le domaine du Traitement Automatique des Langues Naturelles (TALN) et plus précisément dans celui de l'Extraction d'Information. Le but des travaux en Extraction d'Information est de développer des méthodes et des outils visant à extraire automatiquement des informations à partir de textes écrits en langue naturelle. Il s'agit d'analyser des documents textuels afin de collecter et de structurer des informations précises définies en amont. De manière générale, les types d'informations recherchées sont décrits formellement à travers des formulaires dits d'extraction. Cette technologie hérite des travaux en structuration puis en compréhension de textes. Elle a acquis sa maturité lors des années 1990 au cours desquelles ont émergé les premiers véritables systèmes d'Extraction d'Information. Des systèmes plus efficaces ont ensuite été développés en se fondant principalement sur des méthodes d'analyse linguistique et/ou d'apprentissage utilisant des outils et des techniques issues des recherches en TALN.

Les systèmes d'Extraction d'Information existants traitent des corpus composés des textes issus de la presse (journaux, revues économiques, dépêches), de la littérature (livres, essais, actes de publications scientifiques) ou de documents officiels d'institutions ou d'entreprises (rapports d'expertise, bilans financiers), nous avons choisi de traiter les journaux et extraire les information pour faciliter au utilisateur la consultation des journaux et il trouve ce qu'il s'intéresse rapidement.

L'objectif de ce travail est de proposer un système d'extraction d'Information à partir les journaux locaux. Dans ce but, nous avons d'abord avoir conçu et analyser tous les composants nécessaires. Ensuite, nous avons essayé de réaliser un système pour extraire et rechercher les informations dans les journaux à partir des mots clés formulés par l'utilisateur.

A cet effet, et pour atteindre l'objectif cité ci-dessus, nous avons organisé ce manuscrit comme suit :

Le premier chapitre, présente le domaine de l'Extraction d'Information. Nous commençons par le définir avant de le placer dans son contexte en le situant historiquement et vis-à-vis d'autres domaines visant à traiter les informations présentes dans les textes.

Dans le deuxième chapitre nous avons utilisé le langage de modélisation UML pour la spécification des besoins, l'analyse et la conception en s'appuyant sur les

principaux diagrammes structurels et comportementaux tels que les diagrammes de classes, de séquence, et de cas d'utilisation.

Le dernier chapitre est consacré à la partie d'implantation du projet. Il expose les différents outils liés au développement de cette application avec quelques interfaces de présentation.

Conclusion et perspectif :

Le travail effectué dans ce projet fait partie du domaine de l'extraction d'information et de traitement automatique des langues naturelles.

Extraire des informations précises à partir de textes constitue aujourd'hui un enjeu considérable dans de nombreux domaines. C'est particulièrement le cas dans les mondes institutionnel, de l'économie, de l'industrie et du journalisme.

Nous avons donné un aperçu général sur l'extraction d'information à partir des fichiers PDF et des fichiers Html.

Dans le travail réalisé, nous avons conçu et développé un outil de recherche et d'extraction d'information à partir des journaux et des sites web. L'application et grâce aux techniques de recherche et d'extraction met entre les mains de l'utilisateur la possibilité de consulter tous se qui se passent autour de notre monde avec des moyens rapides et efficaces.

Le travail réalisé nous a permis de se familiariser avec le langage de conception UML ainsi qu'avec le langage de développement Visual C#.

Pour la continuité de ce travail nous suggérons comme perspectives la recherche et l'extraction à partir les archives des journaux. Nous suggérons aussi l'insertion des ontologies pour avoir un moyen plus efficace dans la recherche de l'information.

Références bibliographiques

[1] G. Attardi et C. Burrini, The PISAB Question Answering System, Proceedings of the Ninth Text Retrieval Conference (TREC 9), Gaithersburg, USA, NIST Special Publication 500-249, pp. 446-451, 2001.

[2] C. A. Will, Comparing Human and Machine Performance for Natural Language Information Extraction: Results from the TIPSTER Text Evaluation, Proceedings of TIPSTER Text Program (Phase I), Morgan Kaufmann Publishers, San Francisco, pp. 179-194, 1993.

[3] C. A. Will, Comparing Human and Machine Performance for Natural Language Information Extraction: Results for English Microelectronics from the MUC-5 Evaluation. Proceedings of the Fifth Message Understanding Conference (MUC-5), Morgan Kaufman Publishers, pp. 53-67, 1993.

[4] A. Tartier, Evolution terminologique : méthodes d'analyse automatique de corpus diachroniques, Actes de RECITAL 2000 (Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues Naturelles), Lausanne, Suisse, pp. 523-527, 2000.

[5] R. Gaizauskas, An Information Extraction Perspective on Text Mining:Tasks, Technologies and Prototype Applications, Euromap Text mining Seminar, Londres, Grande-Bretagne, 2002.

[6] M. T. Pazienza (réd.), Information Extraction (a multidisciplinary approach to an emerging technology), Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence subseries), LNAI 1299, Springer, pp. 10-27, 1997.

[7] R. Yangarber, R. Grishman, P. Tapanainen et S. Huttunen, Automatic Acquisition of Domain Knowledge for Information Extraction, Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), Saarbrcken, Germany, pp. 940-946, 2000.

[8] Y. Wilks, Information Extraction as a Core Language Technology, Information Extraction (a multidisciplinary approach to an emerging technology), M. T. Pazienza (réd.), Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence subseries), LNAI 1299, Springer, pp. 1-9, 1997.

[9] W. Lehnert, C. Cardie, D. Fisher, J. McCarthy, E. Riloff et S.Soderland (1994), Evaluating an Information Extraction System, Journal of Integrated Computer-Aided Engineering, Vol. 1(6), pp. 453-472, 1994.

[10] C. Cardie, Empirical Methods in Information Extraction, AI Magazine, Vol. 18(4), pp. 65-79, 1997.

[11] S. G. Soderland, D. Fisher, J. Aseltine et W. Lehnert, CRYSTAL: Inducing a Conceptual Dictionary, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95), Montréal, Canada, pp. 1314-1319, 1995.

[12] S. G. Soderland, D. Aronow, D. Fisher, J. Aseltine et W. Lehnert, Machine Learning of text analysis rules for clinical records, Technical Report TE-39, University of Massachusetts, 1995.

[13] B. M. Sundheim (réd.), MUC-5, Proceedings of the Fifth Message Understanding Conference, Baltimore, USA, Morgan Kaufmann Publisher, 1993.

[14] T. Poibeau, Extraction d'Information à base de connaissances hybrides, Thèse de Doctorat (Spécialité Informatique), LIPN, Université Paris-Nord, France, 2002.

[15] R. D. Holowczak et N. R. Adam, Information Extraction based Multiple-Category Document Classification for the Global Legal Information Network, Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence, AAAI/MIT Press, pp. 1013-1018, 1997.

[16] Y. Ichimura, Y. Nakayama, M. Miyoshi, T. Akahane, T. Sekiguchi et Y. Fujiwara, Text Mining System for Analysis of A Salesperson's Daily Report, Proceedings of PACLING'01 (Pacific Association for Computational LINGuistics), Kitakyushu, Japon, 2001.

[17] W.J. Frawley, G. Piatetsky-Shapiro et C.J. Matheus, Knowledge Discovery in Databases: An Overview, Knowledge Discovery in Databases, MIT Press, pp. 1-27, 1991.

[18] U.M. Fayyad, G. Piatetsky-Shapiro et P. Smyth, From Data Mining to Knowledge Discovery: An Overview, Advances in Knowledge Discovery and Data-Mining, AAAI/MIT Press, pp. 1-36, 1996.

[19] M. Rajman et R. Besançon, Text Mining: Natural Language techniques and Text Mining applications, Proceedings of the Seventh IFIP 2.6 Working Conference on Database Semantics (DS-7), Leysin, Suisse, pp. 50-65, 1997.

[20]M. Rajman et R. Besançon, Text Mining — Knowledge extraction from unstructured textual data, Proceedings of the Sixth Conference of International Federation of Classication Societies (IFCS-98), Rome, Italie, pp. 473-480, 1998.

[21] R. Lefébure et G. Venturi, Gestion de la relation client, Eyrolles, ISBN 221211331-5, 2005.

[22] G. Salton et M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill Book Company, 1983.

[23] R. Baeza-Yates et B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley Longman Publishing Co, 1999.

[24] C. J. Van Rijsbergen, Information Retrieval, Butterworths, Londres, ISBN 040870929-4, 1979.

[25] R. Gaizauskas et Y. Wilks, Information Extraction: Beyond Document Retrieval, Computational Linguistics and Chinese Language Processing, Vol. 3, n° 2, pp. 17-60, 1998.

[26] W. Lehnert, The Process of Question Answering — A Computer Simulation of Cognition, Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA, 1978.

Résumé

Le but de ce projet est d'analyser et de réaliser une application spéciale pour la recherche de l'information et de l'extraire par les techniques de balayage (parsing). l'objectif principal de ce projet est de retrouver un outil qui permet aux lecteurs de trouver rapidement aussi que facilement les informations qu'ils veulent atteindre , soit dans les journaux ou dans les sites web.la phase conceptuelle du système a été réalisée par le langage UML et la programmation a été faite par Visual C#.

Mots clés : Recherche, extraction, balayage, journaux, UML, C#.

Abstract

The purpose of this project is to concept and develops a special application for information retrieval and extraction by parsing techniques. The main objective of this project is to find a tool that allows readers to quickly find information as easily as they want to achieve, either in newspapers or sites web. The conceptual phase of the system was carried out by the UML and programming was done by Visual C #.

Keywords: search, extraction, parsing, newspapers, UML, C #.

ملخص

يهدف هذا المشروع إلى تصور و إنجاز برنامج خاص بالبحث عن المعلومة ثم استخراجها عن طريق تقنيات المسح. الهدف الرئيسي وراء هذا المشروع هو تقديم وسيلة للقراء تمكنهم من الوصول السريع والمريح للمعلومات التي يريدون الحصول عليها سواء كانت عن طريق الصحف أو مواقع الويب , المرحلة التصورية للبرنامج أنجزت بلغة *UML* والبرمجة تمت بلغة C#

كلمات مفتاح: بحث، استخراج مسح، صحف، UML، C#