UNIVERSITI PUTRA MALAYSIA

BUILDING A FRENCH STEMMER USING A DICTIONARY OF FRENCH ROOT WORDS

FULAYI IDI

FSKTM 1999 2

BUILDING A FRENCH STEMMER USING A DICTIONARY OF FRENCH ROOT WORDS

FULAYI IDI

MASTER OF SCIENCE
UNIVERSITI PUTRA MALAYSIA

1999

BUILDING A FRENCH STEMMER USING A DICTIONARY OF FRENCH ROOT WORDS

By

FULAYI IDI

Thesis Submitted in Fulfilment of the Requirements for the Degree of Master of Science in the Faculty of Computer Science and Information Technology Universiti Putra Malaysia

July 1999

ACKNOWLEDGEMENTS

All praises should go to ALLAH for keeping me strong enough and giving me the courage and patience to finish this work. I would like to express my earnest appreciation and heartfelt thanks to Dr Hjh Fatimah Ahmad, the chairperson of the supervisory committee, for her patience, constructive guidance, kindness and logistic support throughout my studies. Thanks to her, I have been introduced to this wonderful field of information retrieval, and, mostly, she opened my eyes to the exacting but how gratifying world of research. Despite her academic and administrative obligations, she was able to find enough time for my questions and provided required enlightenment when it mattered most. My sincere gratitude go to Dr Md Yazid Mohd Saman, Puan Norwati, and Encik Muhamad Taufik Abdullah, all members of my supervisory committee, for their constructive suggestions, kind cooperation, patience and encouragement. Their brotherly and sisterly attitudes have been conducive for the completion of my studies.

I also would like to extend special thanks to Dr MD. Nasir Sulaiman, Dr Ramlan Mahmood, Dr Ali B. Mamat, Dr A.K Ramani, Prof Hassan Selamat, Prof Dr Burhanuddin Ali, Encik Rusli bin Haji Abdullah, and Encik Nordin Zakariah, for allowing me to attend their classes. Many thanks also go to Muhamad Othman, PhD student and UPM lecturer, for his continuous encouragement, and to Dr Inuwa Shehu Usman (Nigeria), for his brotherly advice. Special thanks also should go to

Muhammad Djale, Moawia Elfaki, and Sadik Babiker, for their kind assistance during my study. All my friends, Abdullah Seif, Muhammad Hafidh, Muhammad Sulaiman Dadi, Omar Semmasaba, Hamid Efawareh, Nasor Kaboge Mziba, Mussa Maulid Mssamba, all UWATA members at the International Islamic University, all CEPGL members at the International Islamic University, Faiz Elfaki, Nyonya Harun, Malik Idries, Hussein Selemani Kahinga, Lawal Ahmed (Nigeria), Osman Harun (Ghana), and many others I can not list here, to all of them, I humbly say 'thank you'.

Last but not the least, my heartfelt thanks should go to my father, Idi Massud, my mothers, Kanyambo Césarie (Johah) and Kurusumu, my brothers Massoud Idi, Madebari Jean Sauveur Fadhili, Juma Ngaiwa, Ali Idi Mtumwa, Gossaji Idi, Ibrahim Idi, Sande Idi, Gammal Idi, Mussa Idi, Ramadhani Bakari Billy, Gandhi Idi, Mubarak Idi, Amini Idi, and sisters, Mami Idi, Alimah Idi, the late Mayassa Idi, Adija Idi, Rukiya Idi, Zayi idi, Kiboga Idi, and the late Shida Idi, for their prayers, encouragement, and understanding, which have always been a source of inspiration and strength throughout my life up to this moment. Staying abroad for a long period of time would be unpleasant, sometime impossible, without the presence of friends. For all those whose names were not mentioned here, the moral support and friendship they offered will be remembered.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Abstract of the thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the Degree of Master of Science.

BUILDING A FRENCH STEMMER USING A DICTIONARY OF FRENCH ROOT WORDS

By

FULAYI IDI

July 1999

Chairperson : Hjh. Fatimah Ahmad, PhD.

Faculty: Computer Science and Information Technology

In this thesis, a strong French stemming algorithm based on a dictionary of French root words is developed. Four modules are observed for this purpose.

The first module deals with the development of a list of French root words, and a list of affixes, that is, prefixes, suffixes, and prefix-suffix pairs.

The second module removes the punctuation from words to be stemmed. It also removes stop words from the corpus to be stemmed. After this second module, words are noise-free, and this leads to the third module, that is, the stemming proper.

The stemming order adopted is prefix, then suffix, and finally prefix-suffix pairs. Any word to be stemmed is first compared to a dictionary of French root words to check if it is a root word. Then, the actual stemming process is performed.

The stemming algorithm constructed is tested using selected criteria, among which are inflection removal, prefix stripping and suffix stripping. For all these tests, the new French stemming algorithm performs better than the existing French stemmer, Savoy's stemmer.

Tests are also carried out to check the performance of the new French stemmer in terms of understemming, overstemming, ambiguous stemming and dictionary error. The new French stemmer has fewer understemming, overstemming, and ambiguous stemming than Savoy's stemmer. However, the new stemmer has more dictionary born errors than Savoy's stemmer.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk mendapatkan Ijazah Master Sains.

MEMBINA SATU PENGAKAR PERKATAAN BAHASA PERANCIS MENGGUNAKAN KAMUS KATA AKAR BAHASA PERANCIS

Oleh

FULAYI IDI

July 1999

Pengerusi: Hjh. Fatimah Ahmad, Ph.D

Fakulti: Sains Komputer dan Teknologi Maklumat

Dalam tesis ini, satu algoritma pengakar perkataan bahasa Perancis berdasarkan kepada kamus kata akar bahasa Perancis telah dibangunkan. Empat modul telah ditinjau bagi tujuan ini.

Modul pertama melaksanakan pembangunan satu senarai kata akar bahasa Perancis, dan senarai imbuhan iaitu awalan, akhiran dan pasangan awalan-akhiran.

Modul kedua akan menghilangkan tanda bacaan daripada perkataan yang hendak dicari kata akarnya. Ia juga menghilangkan kata henti daripada koleksi dokumen yang hendak dicari kata akarnya. Selepas modul kedua ini, ayat dikatakan

dokumen yang hendak dicari kata akarnya. Selepas modul kedua ini, ayat dikatakan bebas hingar, dan ini diteruskan kepada modul ketiga, iaitu, proses pengakaran sebenar.

Tertib pengakaran perkataan yang digunakan adalah awalan, kemudian akhiran dan akhir sekali pasangan awalan-akhiran. Sebarang perkataan yang akan dicantas mesti dibandingkan dengan kamus kata akar bahasa Perancis dahulu untuk memeriksa sama ada ianya merupakan kata akar atau tidak. Kemudian, barulah proses cantasan dijalankan.

Algoritma pengakaran yang dibina telah diuji menggunakan kriteria terpilih, di antaranya adalah pembuangan fleksi, pencantasan imbuhan awalan dan akhiran. Bagi kesemua pengujian ini, algoritma pengakaran bahasa Perancis yang baru ini menunjukkan prestasi yang lebih baik berbanding dengan pengakar bahasa Perancis yang sedia ada, iaitu pengakar Savoy.

Pengujian juga telah dijalankan untuk memeriksa prestasi pengakar bahasa Perancis yang baru ini dari segi terkurang cantas, terlebih cantas, pengakaran yang meragukan dan ralat kamus. Pengakar bahasa Perancis ini mempunyai kurang ralat daripada aspek terkurang cantas, terlebih cantas, dan pengakaran yang meragukan berbanding dengan pengakar Savoy. Walau bagaimanapun, pengakar baru ini mempunyai lebih banyak ralat kamus berbanding pengakar Savoy.

CHAPTER I

INTRODUCTION

Background Overview

Information retrieval is quite a new discipline in the sense that many of its modern applications depend on concepts that have been formulated only during the past few decades. Nevertheless, the subject has roots that extend back through many centuries.

The traditional library, a collection of documents, led to the development of standard procedures for manual cataloguing, use of card indexes, bibliographies, and the circulation and ordering of books, journals, and reports.

However, the traditional library was oriented more to the provision of documents than the supply of information. This orientation is efficient, provided library users are interested primarily in the well-defined subjects covered by a small number of books and journals.

Yet, at the present time, there are many fields of study whose investigation requires information for a number of different disciplines, and for which requests for relevant information can not be met by reference to a small, easily specified set of documents. In fact, social and technical changes are taking place very rapidly, as the result of numerous discoveries. The consequences of these new discoveries affect lives of entire populations to a degree that has never been the case previously. Thus, more and more people are vitally interested in having fast access to more and more information. The ability to maintain the rapid growth of technological and social changes requires that the vast amount of information be instantly available than required.

Current Information Production Trend

Presently, it can be witnessed that the number of available on-line textual databases increases rapidly. These databases contain a massive variety of information, including full text newspaper stories, stock quotations, airline schedules, journal abstracts, instruction manuals, library catalogs, census data, to name just a few. A number of reasons, including Internet factor and technological discoveries, can be used to explain the now prevailing information production frenzy.

Internet Factors

Regarding the Internet factor, Information highway has introduced an era referred to as information society. In fact, Internet has acquired the status of a vast

library, as well as worldwide communication medium. It is now being used by academicians and researchers to reach information spanning across the wide range of Internet nodes. In addition, the www merges techniques of information retrieval and hypertext to make information circulation easy globally (Hartley, 1995).

Technology Factors

The concept refers to the growing power of workstations, maturity of networking and user interfacing that enable quick spread of information between and within research institutions. For example, electronic data interchange has discarded paper involvement, and researchers can communicate using computer to computer links. Also, primitive forms of interaction have completely been replaced by electronic means of communication. In addition, a host of ideas, including network computing, which allows users to access resources from several different locations without worrying how the information is being accessed, the concepts of client-server, and cooperative processing,..Etc, all are favorable to unhindered production and accumulation of masses of information (Tapscott, 1993). Organising this mass of information in order to make it available for timely consumption by users is the work Information Retrieval.

Information Retrieval

The term information retrieval refers to the activities involved in searching a body of literature in order to find items that deal with a particular subject area that

responds to particular information needs. More specifically, information retrieval achieves text representation, organisation, and carries out the retrieval of stored information items that are relatively similar to information requests received from users. The task of retrieving information is then a complex exercise that involves information needs from a user, a computer, and a database, all of which combine to form an information retrieval system.

An Information Retrieval System

An information Retrieval System is any tool or device that organises a body of literature in such a way that it can be searched conveniently (Lancaste, 1978). An information retrieval system, as shown in Figure1, contains three major components, that is, the input, processor and output. During the first stage, a text document is used in order obtain a representation of each document and query suitable enough to be used by a computer. Secondly, there is the processor, which is a component involved in the structuring of information in some appropriate way, such as representation and classification. Finally, we come to the output, which is usually a set of citations or document numbers. In an attempt to improve information representation and processing, these three information retrieval system components have of recent been subjected to intensive research effort, and the chief area of particular scrutiny has been the improvement of conflation algorithms. The results of research in conflation algorithms have particularly been fruitful for the English language which altogether counts 4 stemming algorithms, Arabic, which counts 4 stemming algorithms,

stemming algorithms, Slovene, which has 2 stemming algorithms, and Malay, which has also 2 conflation algorithms. French comes last with only 1 conflation algorithm.

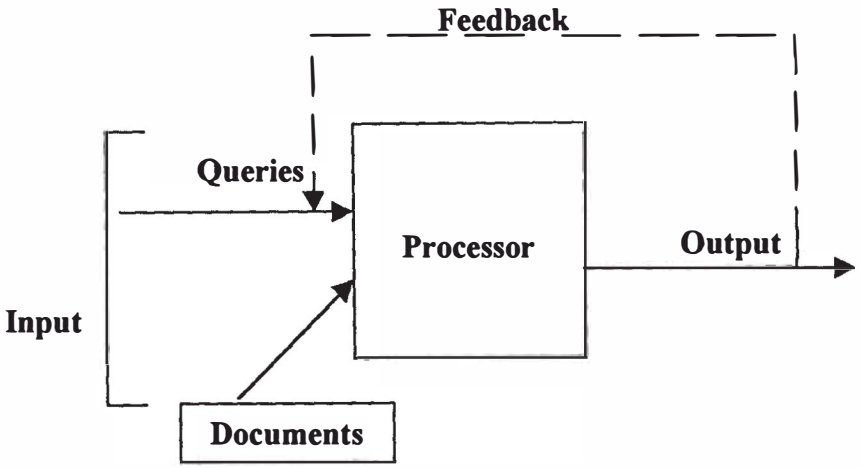

Figure 1.1: Information Retrieval System

Statement of the Problem

Automatic characterisation, in which a software attempts to duplicate the human process of 'reading' is now a reality for many languages. Although vast improvements have been scored towards reaching this goal, none of the existing characterisations, technically known as conflation algorithms, can perform to a maximum 100%. Regarding French especially, the performance of the only existing conflation algorithm, that is, the Savoy's Stemmer, is even lower. Given the extent of error rate for the Savoy's Stemmer, more efforts are needed to look into common errors pointed out by Savoy, especially, rules and spelling adjustments, irregular compound words, and the simultaneous removal of prefixes and suffixes (Savoy, 1993). In addition, the existing French Stemmer does not use a dictionary of root

words, as there is no such document in French. We are intending to build a French stemmer that is based on a dictionary of French root words. This latter dictionary will be designed by transforming an existing French one. Furthermore, the existing French stemmer's performance needs to be improved, so as to make it more reliable. For this purpose, the same data used by Savoy for his stemmer will once again be repeated in order to test the improvement scored by the new French stemmer, this in comparison with the existing Savoy's stemmer.

Objectives of the Study

The major objective of the study is to build a French stemmer using a dictionary of French root words. The following specific objectives can be singled out:

1. To build a new French stemmer for information retrieval purposes;
2. To evaluate the performance of the stemmer.

Contribution of the Study

This piece of research undertakes to build a French stemmer using a dictionary of French root words. This is a new approach, as far as French is concerned, and this on a double basis: first, there exists no dictionary of French root words up to date, and we have tried, in this research, to build one. Secondly, a new stemming approach based on the dictionary of French root words has been built.

Scope of the Study

This work has got some limitations that we have to underline:

. The dictionary of French root words is obtained by changing an existing French dictionary; this first attempt may not be quite representative.

. The testing of the built stemmer will be done on the same data as for the existing Savoy's stemmer; the testing then will not be quite representative to the extent of being standardised.

Structure of the Thesis

Chapter I briefly highlights the historical aspect of the field of Information Retrieval. The chapter also attempts to define the concepts of information retrieval, and information retrieval systems.

Chapter II is all about the review of the literature. In this respect, the concept of word conflation is discussed in detail; also, ample information on the nature of the French grammar, especially word suffixation and word prefixation is provided. The chapter also accounts for some of the existing conflation algorithms that use word stemming, namely the Savoy's Stemmer, Asim's Stemmer, Fatimah's Stemmer, Latin Stemmer, Okapi 86, English and Arabic Stemmers.

As for Chapter III, it deals with the methodology referred to in the course of the project. More specifically, it looks into the architectural design and implementation of the intended French stemmer. This architectural design is conceived in terms of modules, notably the stopword identification removal, lexical parsing, punctuation removal, prefix identification and removal, suffix identification and removal, prefix and suffix identification and removal modules.

Chapter IV discusses about the design and implementation of the stemmer proper. The chapter especially dwells on each and every step that is part of the stemming process and tries to account for the why, what and where of the entire stemming fabric.

Chapter V consists of the results and discussion of the study, whereby the results from the various tests conducted will be pondered over.

As for Chapter VI, it contains the conclusion. In the process, some recommendations will be made for possible future study.

CHAPTER II

LITERATURE REVIEW

A lot of efforts across time, language, and space have been made in an attempt to achieve proper removal of affixes (still today). This chapter reviews some grammatical aspects of the French language, and presents major stemming algorithms that exist in selected languages.

The topics discussed therefore comprise the nature of word conflation, a morphological overview of the French grammar, Malay Stemmers, the Latin Stemmer, Okapi'86, Arabic and English Stemmers. First and foremost, let us get acquainted with the nature of word conflation.

French Language: Morphological Overview

French is one of the most widely spoken languages in the world, and has numerous and more complex inflections than English. In French, nouns, adjectives, pronouns, articles, etc, are declined according to gender (Masculine, feminine), and number (singular, plural); also, person and tense are added to verbs (Savoy, 1993).