

Supporting the Chinese Language in Oracle® Text

Masterarbeit

im Fach Arbeits-, Lern- und Präsentationstechniken
Masterstudiengang Informationswirtschaft
der
Fachhochschule Stuttgart –
Hochschule der Medien

Poh Choo Lai

Erstprüfer: Prof. Dr.-Ing. Peter Lehmann
Zweitprüferin: Frau Barbara Steinhanses

Bearbeitungszeitraum: 20. Okt 2003 bis 31. Mär 2004

Stuttgart, März 2004

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Masterarbeit selbständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht.

Ort, Datum

Unterschrift

Kurzfassung

Gegenstand dieser Arbeit sind die Problematik von chinesischem Information Retrieval (IR) sowie die Faktoren, die die Leistung eines chinesischen IR-System beeinflussen können. Experimente wurden im Rahmen des Bewertungsmodells von „TREC-5 Chinese Track“ und der Nutzung eines großen Korpus von über 160.000 chinesischen Nachrichtenartikeln auf einer *Oracle10g* (Beta Version) Datenbank durchgeführt. Schließlich wurde die Leistung von *Oracle® Text* in einem so genannten „Benchmarking“ Prozess gegenüber den Ergebnissen der Teilnehmer von TREC-5 verglichen. Die Hauptergebnisse dieser Arbeit sind: (a) Die Wirksamkeit eines chinesischen IR Systems ist durch die Art und Weise der Formulierung einer Abfrage stark beeinflusst. Besonders sollte man während der Formulierung einer Anfrage die Vielzahl von Abkürzungen und die regionalen Unterschiede in der chinesischen Sprache, sowie die verschiedenen Transkriptionen der nicht-chinesischen Eigennamen beachten; (b) Stopwords haben keinen Einfluss auf die Leistungsfähigkeit eines chinesischen IR Systems; (c) die Benutzer neigen dazu, kürzere Abfragen zu formulieren, und die Suchergebnisse sind besonders schlecht, wenn Feedback und Expansion von Anfragen („query expansion“) nicht genutzt werden; (d) im Vergleich zu dem *Chinese_Vgram_Lexer*, hat der *Chinese_Lexer* den Vorteil, reale Wörter und einen kleineren Index zu erzeugen, sowie höhere Präzision in den Suchergebnissen zu erzielen; und (e) die Leistung von *Oracle® Text* für chinesisches IR ist vergleichbar mit den Ergebnissen von TREC-5.

Schlagwörter:

Chinesische Sprache, chinesisches Information Retrieval, Leistungsbewertung, Benchmarking, Text Retrieval Conference (TREC), Oracle® Text

Abstract

Issues concerning Chinese information retrieval (IR) and factors influencing retrieval effectiveness in Chinese IR were investigated. Experiments were conducted on an *Oracle10g* (beta version) database using the evaluation framework of the TREC-5 Chinese Track and a large corpus of over 160,000 Chinese news articles. Finally, the performance of *Oracle® Text* in Chinese IR was benchmarked against the results of TREC-5 participants. The main conclusions drawn are: (a) retrieval effectiveness is highly influenced by the ways in which queries are formulated; and it is important during the formulation of a query, to take into consideration the extensive use of abbreviated noun-phrases and the regional differences present in the Chinese language, as well as the different standards of transliteration of non-Chinese proper names; (b) the presence of stopwords does not affect retrieval performance; (c) real users tend to formulate short queries, and retrieval performance is worse when feedback and query expansion are not applied; (d) as compared to the *Chinese_Vgram_Lexer*, the *Chinese_Lexer* has the advantage of generating real word tokens and a smaller index, as well as producing retrieval results of higher precision; and (e) *Oracle® Text*'s retrieval performance for Chinese IR is comparable to that of TREC-5 results.

Keywords:

Chinese language, Chinese information retrieval, retrieval performance evaluation, benchmarking, Text Retrieval Conference (TREC), Oracle® Text

Table of Contents

Erklärung	ii
Kurzfassung	iii
Abstract	iv
Table of Contents	v
List of Figures	viii
List of Tables	viii
1 Introduction	1
1.1 Trends and Issues in Chinese Information Retrieval	1
1.2 Aims and Objectives	1
1.3 Outline of the Thesis	2
2 The Chinese Language	3
2.1 A Brief History of the Development of the Chinese Language	4
2.2 The Written Chinese	6
2.2.1 Classification of Writing Style	6
2.2.2 Character Forms – Traditional versus Simplified	7
2.2.3 Bopomofo (Zhuyin) and Pinyin	8
2.3 Characteristics of Modern Chinese	9
2.3.1 Standardization of Chinese Characters	9
2.3.2 Definition of a Word	10
2.3.3 Variant Words, Homophones and Homographs	12
2.3.4 Political and Cultural Influences	14
2.3.5 Transliteration Differences	15
2.4 Implications on Information Processing and Retrieval	16
3 Chinese Information Processing	17
3.1 Coded Character Set Standards and Encoding Methods	17
3.1.1 Introduction	17
3.1.2 Unicode, ISO 10646 and UTF Encodings	18
3.1.3 National Character Set and Encoding Standards	19
3.2 Input Methods	21
3.2.1 Direct Entry of Character Code Values	21
3.2.2 Zhuyin and Pinyin Input Methods	21
3.2.3 Input by Character Structure	23
3.2.4 Other Input Methods	24
3.3 Chinese Environment Setup	24

3.3.1	Microsoft® Windows®	24
3.3.2	MacOS	26
3.3.3	Unix and Linux	26
4	Indexing Methods for Chinese IR Systems	28
4.1	Character-based Indexing (1-gram Indexing)	28
4.2	N-gram Indexing	29
4.3	Word-based Indexing	30
5	Retrieval Performance Evaluation	31
5.1	Recall and Precision	31
5.1.1	Definitions	31
5.1.2	Recall-Precision Curve and Precision Interpolation	33
5.1.3	Average Precision and R-Precision	35
5.1.4	Precision Histogram	35
5.2	The Text Retrieval Conference (TREC)	36
5.2.1	Tasks and Tracks	36
5.2.2	Relevance Judgment	37
5.2.3	Evaluation Measures	38
5.2.4	Chinese Document Retrieval in TREC-5 and TREC-6	38
5.3	Limitations of TREC Evaluation	40
5.3.1	Relevance Assessments and Missed Documents	40
5.3.2	Pooling and Averaging	40
6	Oracle® Text and Chinese IR	41
6.1	Database Character Sets and NLS Setting	41
6.2	Indexing Chinese Text	43
6.2.1	Context Index	43
6.2.2	Lexer, Wordlist and Stoplist	44
6.3	Querying	46
6.3.1	CONTAINS Operator	46
6.3.2	Scoring Algorithm	47
6.3.3	CONTAINS Query Operators	47
7	Performance Evaluation of Oracle® Text in Chinese IR	51
7.1	Overview	51
7.1.1	Aims and Objectives	51
7.1.2	Resources	52
7.1.3	Experiment Environment	53
7.2	Experiments	54
7.2.1	<u>Experiment I</u> : Performance Comparison between <i>Chinese_Vgram_Lexer</i> and <i>Chinese_Lexer</i>	54
7.2.2	<u>Experiment II</u> : Effect of Weighted Search Terms on Retrieval Performance	59
7.2.3	<u>Experiment III</u> : Effect of Stopwords on Retrieval Performance	63

7.2.4	<u>Experiment IV</u> : Effect of Thesaurus on Retrieval Performance	64
7.2.5	<u>Experiment V</u> : Retrieval Performance for Short Queries	66
7.3	Benchmarking Performance against TREC-5 Results	70
7.3.1	TREC-5 Result Sets	70
7.3.2	Comparative Study – Automatic Run	71
7.3.3	Comparative Study – Manual Run	73
7.3.4	Results Analysis	75
7.3.5	Benchmarking Results	76
7.4	Conclusion	77
8	Conclusions and Recommendations	79
8.1	Issues in Chinese IR	79
8.2	Potential of <i>Oracle® Text</i> in Supporting Chinese IR	79
8.3	Recommendations for Future Research	80
	Appendix A: Topic Statements of TREC-5 Chinese Track	81
	Appendix B: Index Description	85
B.1	<i>ch_index</i>	85
B.2	<i>zh_index</i>	87
	Appendix C: Search Statements (Queries)	90
C.1	Experiment I & III	90
C.2	Experiment II	91
C.3	Experiment IV	92
C.4	Experiment V	93
	Appendix D: Search Results	96
D.1	Experiment I	96
D.2	Experiment II	100
D.3	Experiment III	102
D.4	Experiment IV	104
D.5	Experiment V	106
	Appendix E: Stoplists	112
E.1	<i>Oracle's</i> Default Stoplist for Simplified Chinese (stop2)	112
E.2	<i>Oracle's</i> Default Stoplist for Traditional Chinese	112
E.3	Modified Stoplist (stop3)	112
	References	113
	Acknowledgments	117

List of Figures

Figure 2.1: Chinese Dialects Spoken in Different Parts of Mainland China	3
Figure 2.2: Chinese Language Tree	4
Figure 3.1: Zhuyin Keyboard Array	22
Figure 3.2: Enabling Chinese Support in Microsoft® Windows® Operating Systems ..	25
Figure 3.3: Selecting an Input Language from the “Language Bar” from a German Operating System.....	26
Figure 5.1: Recall and Precision for a Given Example Information Request.	31
Figure 5.2: Trade-off Between Recall and Precision.....	32
Figure 5.3: Computation of Recall-Precision Points (non-interpolated)	33
Figure 5.4: Recall-Precision Curve (interpolated and non-interpolated)	34
Figure 5.5: A Precision Histogram for Ten Hypothetical Queries.....	36
Figure 5.6: Sample Topic Statement for <i>Ad-hoc</i> Retrieval in the Chinese language. ...	37
Figure 6.1: NLS Architecture for Storing, Processing and Retrieval of Data in Simplified and Traditional Chinese	42
Figure 6.2: Hybrid Algorithm of <i>Chinese_Lexer</i>	45
Figure 7.1: Percentage Increase in Precision Values Between “ch_manual” And “zh_manual” Across All Topics	58

List of Tables

Table 2.1: Classification of Chinese Characters	5
Table 2.2: A Comparison of Classical and Vernacular Chinese Writing Styles	6
Table 2.3: Example of Cantonese Writing Style.....	7
Table 2.4: Chinese Speaking Countries and Their Writing Systems	7
Table 2.5: One-to-One Mapping between the Pinyin and Zhuyin Systems	8
Table 2.6: Standardized Chinese Character Lists in Taiwan	9
Table 2.7: Examples of Morphological Types in the Chinese Language	11
Table 2.8: Examples of Homographs in Modern Chinese.....	13
Table 2.9: Examples of Words Defined Differently in Mainland China & Singapore.....	15
Table 2.10: A Comparison of Transliterations of Non-Chinese Proper Names.....	15
Table 3.1: Ideographs Not Unified	18
Table 3.2: Ideographs Unified.....	19
Table 3.3: Pinyin Input Method for Single Characters and Compound Words.....	22
Table 3.4: Pinyin Input Examples.....	23
Table 3.5: Input Methods by (a) Character Structure, and (b) Pronunciation and Character Structure	24
Table 3.6: Chinese Localized Versions of Microsoft Windows® 2000, NT and XP	25
Table 5.1: A Comparison of Average Precision and R-precision	35
Table 5.2: Indexing and Retrieval Approaches Used in the Chinese Track of TREC-5 and TREC-6.....	39

Table 6.1: <i>Oracle</i> -supported Character Sets for Chinese Characters	42
Table 6.2: NLS_LANG Settings for Different Chinese Locales.....	43
Table 6.3: Preference Classes (adapted from <i>Oracle Corp.</i> (2003))	44
Table 6.4: Categorization of Query Operators	47
Table 7.1: Revision of the SPH Thesaurus for Use in <i>Oracle® Text</i>	52
Table 7.2: A Comparison of <i>zh_index</i> and <i>ch_index</i>	54
Table 7.3: Summary of Search Results for Experiment I	55
Table 7.4: Summary of Search Results for Experiment II	60
Table 7.5: Comparison of Average Precision & R-precision Values in Experiment II ...	62
Table 7.6: Run and Stoplist specification for the six sets of readings	63
Table 7.7: Summary of Search Results for Experiment IV.....	64
Table 7.8: Summary of Search Results for Experiment V.....	67
Table 7.9: Number of Queries Containing Words with Six or More Characters	68
Table 7.10: Distribution of Best Precision Values	70
Table 7.11: Total Number of Result Sets for the Automatic and Manual Runs in the TREC-5 Chinese Track	71
Table 7.12: Summary of TREC-5 Search Results (Automatic)	71
Table 7.13: TREC-5 Recall-Precision Values (Automatic).....	72
Table 7.14: TREC-5 Document Level Precision Values (Automatic)	73
Table 7.15: Summary of TREC-5 Search Results (Manual Run).....	73
Table 7.16: TREC-5 Recall-Precision Values (Manual).....	74
Table 7.17: TREC-5 Document Level Precision Values (Manual)	75
Table 7.18: Ranking of <i>Oracle® Text</i> performance against TREC-5 Results	76

1 Introduction

1.1 Trends and Issues in Chinese Information Retrieval

With the advent of Internet technologies and the increasing amount of Chinese documents that is available on the World Wide Web (WWW), substantial interest has been attracted in the field of Chinese Information Retrieval (IR). Both *Yahoo!* and *Google*, for example, provide different Chinese localized versions of their search engines to serve the Chinese speaking communities; and the annual *Text Retrieval Conference* (TREC) co-sponsored by the *National Institute of Standards and Technology* (NIST) and the *Defense Advanced Research Projects Agency* (DAPRA) has also conducted comparative evaluation on Chinese IR systems.

The uniqueness of Chinese IR is fundamentally related to the fact that Chinese is an ideographic language, whereby sentences are formed by continuous strings of characters, without any distinct indication of word boundaries. Hence, special indexing methods different from those of most Indo-European languages need to be used. In addition, the Chinese language uses a repertoire of thousands of characters which can vary in time and over regions – for example, characters used in ancient Chinese versus characters in modern Chinese; simplified Chinese characters used in Mainland China versus traditional Chinese characters in Taiwan; and special characters for the Cantonese dialect that are used in Hong Kong but not in Taiwan and Mainland China, etc.

The processing of Chinese characters on computer systems is also far from being standardized. Different standards of character sets and encoding methods have been developed to meet the needs of the individual countries or regions; and many proprietary standards developed by software companies are also being widely used today.

1.2 Aims and Objectives

This thesis studies some issues concerning Chinese IR, and investigates factors that affect retrieval effectiveness in Chinese IR.

The specific objectives of this thesis are:

- (i) To provide an overview on the characteristics of the Chinese language, and to discuss how these characteristics may affect Chinese IR
- (ii) To summarize the current approaches to Chinese information processing, namely character set standards, character encoding standards and input methods

- (iii) To study the TREC evaluation framework, and to provide an overview of the IR approaches that have been used in the Chinese tracks of TREC-5 and TREC-6
- (iv) To customize an *Oracle10g* (beta version) database to perform Chinese IR on a large collection of newspaper articles drawn from the *People's Daily* and *Xinhua News Agency*
- (v) To investigate factors that affect retrieval effectiveness in Chinese IR, through the use of the TREC evaluation framework and *Oracle® Text*
- (vi) To benchmark the performance of *Oracle® Text* in Chinese IR

1.3 Outline of the Thesis

This thesis combines a theoretical approach to the study of issues concerning Chinese IR, and an experimental approach to the investigation of factors affecting retrieval effectiveness of a Chinese IR system.

Following this introductory chapter, the characteristics of the Chinese language will be discussed in Chapter 2, and in Chapter 3, the processing of Chinese characters on computers will be covered. Then an overview of the different indexing approaches for Chinese IR, namely character-based indexing, n-gram indexing and word-based indexing, is provided in Chapter 4, followed by an analysis of the TREC-5 evaluation model in Chapter 5 that comprises also a study of the two commonly used evaluation measures, namely recall and precision. In Chapter 6, the basics of indexing and text search in *Oracle® Text* will be covered, including examples that are based on the Chinese language. The performance of *Oracle® Text* for Chinese IR is evaluated in Chapter 7, where the observations and results of a total of five sets of experiments conducted to investigate factors affecting retrieval effectiveness, as well as the overall benchmarking results are reported. Finally, Chapter 8 draws together conclusions from all findings in this thesis, and makes recommendations for future research in Chinese IR and *Oracle® Text*.

2 The Chinese Language

The Chinese Language (汉语 huànyǔ, 华语 huáyǔ, or 中文 zhōngwén) is the most widely used language in the world. The spoken form of Chinese includes dialects that are spoken in different parts of Mainland China (People's Republic of China). Although each of these dialects is spoken differently, they are mostly unified through common writing systems.



Figure 2.1: Chinese Dialects Spoken in Different Parts of Mainland China (adapted from Microsoft® Corp. (2002): Encarta Enzyklopädie)

Mandarin is the official language of Mainland China and Taiwan (Republic of China), as well as one of the four official languages of Singapore, and one of the six languages of the United Nations. Other Chinese dialects, especially Cantonese (Yue) and Hokkien (Min), are often spoken by Chinese immigrants in the Southeast Asia. In Hong Kong, the language of education and formal speech is Cantonese. However, since Hong Kong's reunification with Mainland China in 1997, a rising trend in the use of Mandarin has been observed.

This Chapter shall provide some background about the Chinese language, and discuss its implications on text processing and information retrieval.

2.1 A Brief History of the Development of the Chinese Language

The Chinese language has a long history of over 3,000 years. Chinese characters or hànzi (汉字) are believed to have originated in the Shang (商 Shāng) dynasty between 16th and 11th century B.C., as pictograms depicting concrete objects. These archaic Chinese pictograms were inscribed on tortoise shells and flat cattle bones, and are therefore commonly referred to as the “Oracle Bones” characters (甲骨文 jiǎgǔwén).

The Zhou (周) dynasty has also played an important role in developing the Chinese language during its 900 years of regime (1027 to 221 B.C). Thousands of new characters were created during this period, and these included as well, symbols that were developed mainly on the basis of phonetic similarities between words. The “Philosophy of Taoism” 《道德经》 and the “Five Confucian Classics” 《五经》¹, which are considered important works of the Chinese philosophies, were both written during the *Spring-Autumn Period* and early *Warring States Period* of the Zhou dynasty². The Chinese dialects spoken today are also known to have evolved from the ancient Chinese of the Zhou Dynasty.

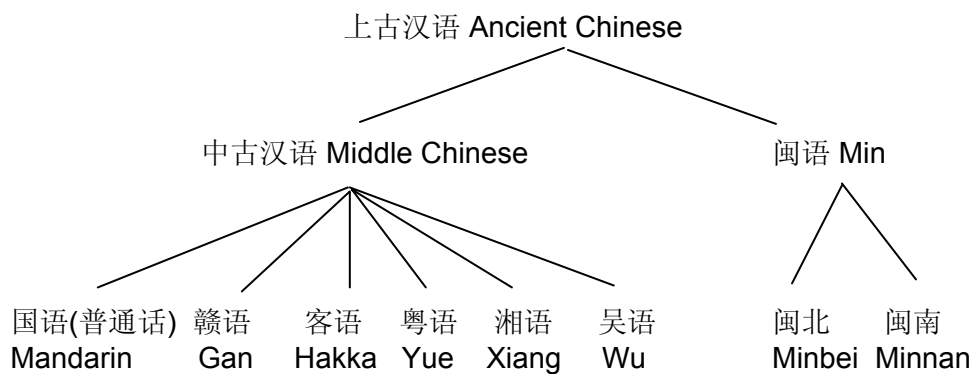


Figure 2.2: Chinese Language Tree (adapted from Campbell, J. (2003): Chinese Dialects, URL: <http://www.glossika.com/en/dict/index.htm>)

¹ The five classics are: *Yijing* 《易经》, *Shijing* 《诗经》, *Shujing* 《书经》, *Zhoujing* 《周礼》 and *Chunqiu* 《春秋》

² The Zhou era was subdivided into three periods:

1027-771 B.C. Western Zhou (西周 xīzhōu)

770-476 B.C. Spring-Autumn Period (春秋时代 chūnqiūshídài)

475-221 B.C. Warring States Period (战国时代 zhànguóshídài)

Until the 18th century, more than 40,000³ characters were recorded in the Chinese language, and today, the total number of Chinese characters is estimated to be around 85,000. Broadly speaking, all Chinese characters can be classified into four classes: (a) *pictograms* that depict concrete objects; (b) *ideograms* that depict abstract concepts; (c) *radical-radical compounds*, in which each element (radical) of a character hints at the meaning; and (d) *radical-phonetic compounds*, in which one component (the radical) indicates the kind of concept the character describes, and the other hints at the pronunciation. Below are some examples of the different classes of Chinese characters:

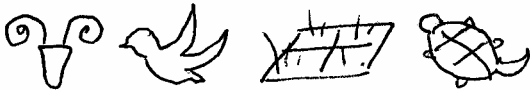
Character Class	Examples
Pictograms	 <p>From left to right: 羊 (sheep), 飞 (fly), 田 (field), 龟 (tortoise)</p>
Ideograms	<p>上 (up) and 下 (down) 一 (one), 二 (two), and 三 (three) 炎 (hot) = 火 (fire) + 火 (fire) 林 (woods or forest) = 木 (wood) + 木 (wood)</p>
Radical-Radical Compounds	<p>拿 (take) = 合 (join together) + 手 (hands) 话 (talk) = 讠 (or 言, which means “language”) + 舌 (tongue) 泪 (tears) = 氵 (symbolizes three drops of water) + 目 (eye) 吠 (bark) = 口 (mouth) + 犬 (dog)</p>
Radical-Phonetic Compounds	<p>妈 (mā “mother”) = 女 (female or woman) + 马 (mǎ “horse”) 摸 (mō “touch”) = 扌 (symbol for “hand”) + 莫 (mò “don’t”) 蚁 (yǐ “ant”) = 虫 (worm) + 义 (yì “justice”) 蕾 (lěi “flower bud”) = 艹 (symbol for “plants”) + 雷 (léi “thunder”) 帽 (mào “hat”) = 巾 (cloth for wrapping) + 冒 (mào “send, give off”)</p>

Table 2.1: Classification of Chinese Characters

Today, some use of Chinese characters is also seen in Korean, Japanese and Vietnamese – referred to as *hanja*, *hanji* and *chữ Hán*, respectively. This is a result of cultural contacts and borrowing of words during the historic times, and these three languages are, however, not related to the Chinese language (Lunde (1999)).

³ The “Kāngxī Dictionary” 《康熙字典》 which was published in 1716, listed a total of 47,021 characters.

2.2 The Written Chinese

2.2.1 Classification of Writing Style

Chinese writings are usually classified into the following three styles:

- (a) Vernacular or Modern Chinese (白话 báihuà)
- (b) Classical Chinese (文言 wényán or 古汉语 gǔhànyǔ)
- (c) Colloquial Chinese (地方话 dìfānghuà)

Vernacular Chinese is the present-day writing style, and is written close to the way which Chinese is spoken. Classical Chinese, on the other hand, is a formal way of writing that is used in classical literature. This writing is a more rhythmic and compact (i.e. using less words) in style, and is still sometimes used today, especially for writing ceremonial documents. Table 2.2 illustrates some differences between the writing styles of Classical Chinese and Vernacular Chinese.

<i>Classical Chinese</i>	<i>Vernacular Chinese</i>	<i>English Explanation</i>
吾所言，合公道否？	我这么说，合理吗？	Is it fair for me to say so?
吾观吕布非常人也。吾若得此人，何虑天下哉！	我觉得观吕这人很有才能。如果有他帮忙，我就不用担心了。	I find Lubu very capable. I should have no worries, if he helped me.
久不相见，今居何处？	好久不见，你近况如何？	We haven't seen each other for a long time. How are you?
兄赐此龙驹，将何以为报？	你送我如此骏马，我该怎么报答你呢？	How should I thank you for giving me such a good horse?
我等为好而来，何乃如此相待。	我们出至好意而来，你怎么如此对待我们呢？	We come here for good will. Why are you treating us like this?

Table 2.2: A Comparison of Classical and Vernacular Chinese Writing Styles (Source of Classical Chinese text: "Romance of the Three Kingdoms" 《三国演义》)

Colloquial Chinese, for example Cantonese, is similar to that of Vernacular Chinese, but it involves the use of characters, words or expressions peculiar to the dialect, such as those underlined in Table 2.3.

<i>Cantonese</i>	<i>English Explanation</i>
國語同埋粵語各有特色，不過如果將國語嘅詩詞換成粵語嚟講，將會出現戲劇性效果。	Both Mandarin and Cantonese possess characteristics of its own. Rewriting Mandarin poems with Cantonese words will bring about drastic effects.

Table 2.3: Example of Cantonese Writing Style

2.2.2 Character Forms – Traditional versus Simplified

Mandarin is the so-called plain Chinese language spoken amongst the Chinese communities. It is referred to as “the common language” (普通话 pǔtōnghuà) in Mainland China, “the national language” (國語 guóyǔ) in Taiwan, and “the Chinese language” (华语 huáyǔ) amongst the overseas Chinese communities, particularly in Southeast Asia. Although Mandarin is spoken the same in these countries or regions, its writing system may vary. For example, the writing system in Taiwan’s Mandarin is based on Chinese characters as they were traditionally written for centuries, and thus Taiwan’s character set is called “Traditional Chinese” (繁体 fántǐ). Mainland China, on the other, uses the “Simplified Chinese” (简体 jiǎntǐ) characters, as a result of its massive writing reforms in the 1950s to promote literacy. Basically, these characters are simplified ideographs of the traditional Chinese characters, and contain fewer strokes. For example, the character “country” (also a word in this case) is written as 国 in simplified Chinese, as compared to 國 in traditional Chinese. Besides, simplified Chinese characters are sometimes created by means of character conflation, that is, by collapsing a few complicated characters into one simpler character. An example is the conflation of 隻 (one piece) and 祇 (only) to 只.

The written form of Cantonese, as used in Hong Kong and Macau, adopts the “Traditional Chinese” characters, but includes also some characters peculiar to the dialect, such as 嚟, 啲, 嘢, 啲, 啲, 啲, etc., as already demonstrated in Table 2.3.

Below is an overview of the different character forms adopted by the various Chinese speaking countries and regions.

<i>Country/Region</i>	<i>Official Dialect</i>	<i>Character Forms</i>
Mainland China	Mandarin	Simplified Chinese
Taiwan	Mandarin	Traditional Chinese
Hong Kong, Macau	Cantonese	Traditional Chinese
Singapore	Mandarin	Simplified Chinese
Malaysia	Mandarin	Traditional and Simplified Chinese

Table 2.4: Chinese Speaking Countries and Their Writing Systems

2.2.3 Bopomofo (Zhuyin) and Pinyin

Bopomofo (ㄅㄆㄇㄏ) or 注音符号 (zhùyīn fúhào “phonetic symbols”) is a set of 37 special phonetic symbols or characters developed in the early 1900s to represent the sounds of spoken Mandarin. The fundamental purpose of Bopomofo, which is still used in Taiwan today, is to teach proper pronunciation to school children.

Pinyin or Hanyu Pinyin (汉语拼音 hànyǔ pīnyīn), which was created by Mainland China in 1958, is a system of romanization for Mandarin Chinese. Since 1979, the International Organization of Standards (ISO) has also adopted the Pinyin system to standardize the transliterations for Chinese proper names⁴.

Although different symbols are used, both Bopomofo and Pinyin are based on the same Mandarin pronunciations, and both systems are made up of 21 consonants (声母 shēngmǔ) and 16 basic vowels (韵母 yùnmǔ)⁵.

<i>Consonants</i>					
b (ㄅ)	d (ㄉ)	g (ㄍ)	j (ㄐ)	zh (ㄓ)	z (ㄗ)
p (ㄆ)	t (ㄊ)	k (ㄎ)	q (ㄑ)	ch (ㄔ)	c (ㄘ)
m (ㄇ)	n (ㄋ)	h (ㄏ)	x (ㄒ)	sh (ㄕ)	s (ㄙ)
f (ㄈ)	l (ㄌ)			r (ㄖ)	
<i>Basic Vowels</i>					
a (ㄚ)	ai (ㄞ)	an (ㄢ)	er (ㄝ)		
o (ㄛ)	ei (ㄟ)	en (ㄣ)	i (ㄝ)		
e (ㄜ)	ao (ㄞ)	ang (ㄤ)	u (ㄨ)		
ê (ㄝ)	ou (ㄛ)	eng (ㄥ)	ü (ㄩ)		

Table 2.5: One-to-One Mapping between the Pinyin and Zhuyin Systems

⁴ Chinese provinces or cities previously known as Peking, Canton and Fukien, etc. are transliterated as Beijing, Guangzhou, and Fujian, following the Pinyin romanization system.

⁵ These 16 basic vowels can be further combined with one another to form compound vowels, e.g. “iang” from “i”+“ang”, “ui” from “u”+“ei”, etc.

2.3 Characteristics of Modern Chinese

2.3.1 Standardization of Chinese Characters

Non-coded Chinese character set standards first appeared in the 1980s in Mainland China and Taiwan for pedagogical purposes, and as an attempt to limit the number of Chinese characters for common use (Lunde (1999)). The Chinese government published a “Table of Modern Chinese Characters for Common Use” (现代汉语通用字表 *xiàndài hànyǔ tōngyòngzì biǎo*) in 1988, which comprises a standardized list of about 7,000 simplified Chinese characters. Within this list are two sub lists that were defined for teaching purposes: (a) “Table of the Most Commonly Used Characters in Modern Chinese” (现代汉语常用字表 *xiàndài hànyǔ chángyòngzì biǎo*) – a list of 2,500 characters to be taught during primary school, and (b) “Table of the Second Most Commonly Used Characters in Modern Chinese” (现代汉语次常用字表 *xiàndài hànyǔ cìyòngzì biǎo*) – another 1,000-character list to be taught in the middle school. In addition, the Chinese government has published a document, entitled “Simplified Character Table” (简化字总表 *jiǎnhuàzì biǎo*), that enumerates 2,249 simplified Chinese characters, and illustrates the traditional forms from which they were derived.

In Taiwan, a total of 43,238 Chinese characters has been standardized for use. These characters are enumerated in four separate lists, as shown in Table 2.6.

<i>List</i>	<i>Pub. Year</i>	<i>Nr. of Char.</i>
Standardized Table of Commonly Used Chinese Characters (常用國字標準字体表 <i>chángyòng guózi biāozhǔn zìtǐ biǎo</i>)	1982	4,898
Standardized Table of Second Commonly Used Chinese Characters (次常用國字標準字体表 <i>cìchángyòng guózi biāozhǔn zìtǐ biǎo</i>)	1982	6,341
Table of Chinese Characters in Rare Use (罕用字体表 <i>hànyòng zìtǐ biǎo</i>)	1983	18,480
Table of Chinese Character Variants (异体國字字体表 <i>yìtǐ guózi zìbiǎo</i>)	1984	18,609
	Total	43,238

Table 2.6: Standardized Chinese Character Lists in Taiwan

2.3.2 Definition of a Word

Unlike many Indo-European languages, the definition of a word in the Chinese language is not a clear and intuitive notion. This is due to the fact a sentence in the Chinese language is written as a continuous string of characters, without any spaces or delimiters between words.

Wang (2002a) defined words (词 *cí*) as the smallest possible standalone units in a sentence, and explained that a word is formed by one or more morphemes (语素 *yǔsù* or 词素 *císù*), from which the semantics of a word is derived. Take for example, the word 有 (*yǒu* or “have”) is made up of just one morpheme, and the word 矿湖 (“quarry”) is composed of two morphemes, namely 矿 (*kuàng* or “minerals”) and 湖 (*hú* or “lake”), and has its meaning derived from the two morphemes. All morphemes named in these examples are also regarded as characters (字 *zì* or 汉字 *hànzì*) for the Chinese speakers. In this sense, one can assume that both morphemes and characters are synonymous in the Chinese language. However, there exist disyllabic or polysyllabic morphemes such as 东西 (*dōngxī* “things”) or 巴基斯坦 (*bājīstǎn* “Pakistan”) (Sproat (2001)) which contradict the assumption.

In any case, studies have shown that 60-70% of all modern Chinese words are disyllabic (i.e. composed of two characters, regardless of whether they are formed by two separate monosyllabic morphemes or just one disyllabic morpheme) (Su (2000b); Wang (2002a); Kua (1993)), as compared to the predominantly monosyllabic words in ancient Chinese.

The ways which words are formed in modern Chinese can be broadly classified as (a) Reduplication, (b) Affixation, (c) Compounding, (d) Proper Names, and (e) Abbreviation, as shown in the examples of Table 2.7.

Morphological Type	Examples
Reduplication	<p>AA: 说 → 说说 (to tell)</p> <p>AAB: 双手 → 双双手 (pairs of hands)</p> <p>AABB: 舒服 → 舒舒服服 (comfortable)</p> <p>A—A: 看 → 看一看 (take a look)</p> <p>ABAB: 研究 → 研究研究 (research)</p> <p>ABB: 亮晶 → 亮晶晶 (sparkling)</p>
Affixation	<p>Prefixation:</p> <p>老 (old) → 老王 (old Wang)</p> <p>小 (small) → 小强 (little Qiang)</p> <p>第 (ordinal prefix) → 第一 (number one)</p> <p>初 (calendrical prefix) → 初三 (3rd day of the lunar month)</p> <p>Suffixation:</p> <p>可 (-able): 可 + 爱 (love) → 可爱 (loveable, cute)</p> <p>可 + 靠 (lean against) → 可靠 (reliable)</p> <p>Diminutive suffixes (like in “duck + ling = duckling”):</p> <p>儿 → 鸟儿 (little bird)</p> <p>子 → 猴子 (monkey)</p> <p>头 → 馒头 (steamed bun)</p> <p>Other derivational suffixes:</p> <p>学 (study): 心理 (mentality) + 学 → 心理学 (psychology)</p> <p>家 (specialist): 物理学 (physics) + 家 → 物理学家 (physicist)</p> <p>化 (-tify): 美 (beautiful) + 化 → 美化 (beautify)</p> <p>率 (rate): 生育 (give birth) + 率 → 生育率 (birth rate)</p> <p>主意 (idea): 马克思 (Karl Marx)+主意 → 马克思主意 (Marxism)</p> <p>了 (already): 吃 (eat) + 了 → 吃了 (have eaten)</p> <p>们 (show plural form): 孩子 (child) + 们 → 孩子们 (children)</p>
Compounding	<p>Root Compounding:</p> <p>蚂蚁 (ant): 蚁王 (ant + queen = queen ant)</p> <p>工蚁 (work+ ant = worker ant)</p> <p>蘑菇 (mushroom): 菇伞 (mushroom + umbrella = pileus)</p> <p>金菇 (gold + mushroom = golden mushroom)</p> <p>Resultative Compounding:</p> <p>Result: 打破 (hit + broken = hit broken)</p> <p>拉开 (pull + open = pulled open)</p> <p>Achievement: 写清楚 (write + clear = write clearly)</p> <p>买到 (buy + arrive = succeeded in buying)</p> <p>Direction: 跳过去 (jump + over there = jump across)</p> <p>走进来 (walk + come in = walk in)</p> <p>Parallel Verb Compounding:</p> <p>购买 (buy) = 购 (buy) + 买 (buy)</p> <p>建筑 (build, building) = 建 (build) + 筑 (build)</p> <p>检查 (examine) = 检 (examine) + 查 (examine)</p> <p>治疗 (cure, treatment) = 治 (cure) + 疗 (cure)</p>

Table 2.7: Examples of Morphological Types in the Chinese Language (Packard (2000); Sproat (2001))

Compounding (cont'd)	<p>Subject-Verb Compounding: 头痛(headache) = 头 (head) + 痛 (hurt) 嘴硬(stubborn) = 嘴 (mouth) + 硬 (hard) 眼红(covet) = 眼 (eye) + 红 (red) 心酸(sad) = 心 (heart) + 酸 (sour) 命苦(tough straits) = 命 (life) + 苦 (bitter)</p> <p>Verb-Object Compounding: 出版(publish) = 出 (emit) + 版(edition) 睡觉(sleep) = 睡 (sleep) + 觉 (sleep) 毕业(graduate) = 毕 (close) + 业 (work, business) 开刀(operate) = 开 (open) + 刀 (knife) 开玩笑(make fun of) = 开 (open) + 玩笑(joke) 照相(take a photo) = 照 (shine) + 相 (image)</p>
Transliteration of Proper Names	<p>尼亚 (níyǎ) for “nia” : 波斯尼亚 (Bosnia), 爱沙尼亚 (Estonia), prefixes and suffixes 斯洛文尼亚 (Slovenia), 罗马尼亚 (Romania), 尼亚加拉瀑布 (Niagara Falls), etc.</p> <p>西亚 (xīyǎ) for “-sia” : 马来西亚 (Malaysia), 突尼西亚 (Tunisia), and “-cia” suffixes 密克罗尼西亚 (Micronesia), 莱蒂西亚(Leticia, Shahani-Ramos), 加西亚·马奎斯 (García Marque, Gabriel), etc.</p>
Abbreviations	<p>亚洲乒乓球联盟 (Asian Association of Table-Tennis) → 亚乒联 工业研究院 (Industrial Research Institute) → 工研院 以色列巴基斯坦和谈 (Israel-Palestinian peace talks) → 以巴和谈 台湾香港 (Taiwan and Hong Kong) → 台港 中国石油 (China Petroleum) → 中油 上海杭州铁路 (Shanghai-Hangzhou Railway) → 沪杭铁路</p>

Table 2.7 (cont'd)

2.3.3 Variant Words, Homophones and Homographs

(a) Variant Words (异形词 yìxíngcí)

In the Chinese language, there exist words that have the same pronunciation and meaning, but are written slightly differently. Here are some examples based on the simplified Chinese characters:

- 狡猾, 狡滑 : Both words are pronounced as “jiǎohuá”, and mean “cunning”
 撤销, 撤消 : Both are pronounced as “chèxiāo”, and mean “cancel”
 耿直, 梗直, 鲠直 : All three are pronounced as “gěngzhí” and mean “straight, or honest”

It has been estimated that there are over 1000 groups of such words, from which 312 groups are single character words, 705 are 2-character words, and 85 are words with more than two characters (Wang (2002a)).

(b) Homophones (同音词 tóngyīncí)

Homophones, which occur most frequently in the Chinese language, are words sharing the same pronunciation but are written differently. Although homophones do not pose any problem in understanding when read, they may cause confusion in listening. For example,

- 这是一个工事 (gōngshì) “This is a fortification”
 这是一个公式 (gōngshì) “This is a formula”
- 这种药物可以治癌 (zhìái) “This type of medicine can cure cancer”
 这种药物可以致癌 (zhìái) “This type of medicine can cause cancer”

(c) Homographs (同形词 tóngxíngcí)

Homographs are words that are written exactly the same, but differ in meaning or pronunciation. A study on 640 groups (a total of 1302 words) of simplified Chinese homographs selected from the “Modern Chinese Dictionary” 《现代汉语词典》 (Su (2000a)) showed that 40% of these homographs are homonyms (同形异义词 tóngxíng yìyì cí, or 同形多义词 tóngxíng duōyì cí) – meaning words that are written and pronounced the same, but have different semantic meanings. Another 42% are related words (引申词 yǐnshēncí), and the remaining 18%, referred to as 体用同称 (tǐyòng tóngchēn), are words that share the same origin, but belong to different parts of speech (POS), such as noun, verb, adjective, etc. Below is an illustration of the different types of homographs present in the Chinese language.

<i>Homonyms</i>	<i>Homographs as Related Words</i>	<i>Homographs with the same origin but different POS</i>
盘缠 (pánchán) to mean “surrounding”, or 盘缠 (pánchan) to mean “travel expenses”.	褒贬 (bāobiǎn) to mean “pass judgment”, or 褒贬 (bāobian) to mean “speak ill of, or blame”	赤膊 (chìbó) to mean either “baring a shoulder” as a verb, or “a bared shoulder” as a noun.
安心 (ānxīn) to mean either “be content, not desiring to make a change”, or “bad intentions”.	宾服 (bīnfú) to mean either “obey”, or “respect”.	出品 (chūpǐn) to mean either “produce” as a verb, or “product” as a noun.

Table 2.8: Examples of Homographs in Modern Chinese (adapted from Su (2000a))

2.3.4 Political and Cultural Influences

As mentioned earlier, the word “Mandarin” is expressed in three different ways across Mainland China, Taiwan and Singapore. These variations could be traced to the differences in political interests: the Chinese government, for example, imposed Mandarin as the “common language” (普通话 pǔtōnghuà) for its folks to promote literacy, and to facilitate communication within the country; Taiwan, on the other hand, uses Mandarin as its “national language” (国语 guóyǔ); and Singapore is believed to have chosen a more neutral term, “the Chinese language” (华语 huáyǔ), in order to distance itself from Mainland China (Wang (2002a)).

A strong ethnic influence on Chinese language can also be seen in the multicultural society of Singapore, whereby the Malays form about 15% of the country’s population, and over 70% of the Singaporeans are Chinese immigrants from the southern provinces of Mainland China. Words of Malay origins such as 甘榜 (gānbǎng “kampong”), 奎笼 (kuílong “kelong”), 巴刹 (bāshā “basar”), etc., as well as Cantonese or Hokkien words like 爽 (shuǎng “feeling good”) and 大耳窿 (dàěrlong “loan sharks”) belong to the everyday vocabulary of the Chinese native speakers in Singapore. Some of these words have been even standardized by the Singapore government for teaching in schools.

Zhou/Xiao (1998) and Wang (2002a, 2002b) have conducted extensive comparisons between the Chinese vocabularies used in Singapore and Mainland China. Their findings showed that besides some minor differences caused by different candidate-choices in word compounding (e.g. 招生 (zhāoshēng) vs. 收生 (shōushēng) for “students recruitment”; 节水 (jiéshuǐ) vs. 省水 (shéngshuǐ) for “save water”)⁶, the major differences were a result of political and cultural influences. Here are some examples:

(a) Local-specific words that are used in Singapore but not Mainland China:

For example, the special words used in the Singapore traffic systems, namely, 易通卡 (“EZ-Link”, an electronic stored value fare card), 电子乘车指南 (electronic bus guide), 卫星传召德士系统 (satellite taxis tracking system), and 高速公路监察与提示系统 (expressway monitoring advisory system), 用车证 (certificate of entitlement), etc.

⁶ Both 招生 and 收生 are compound words formed from 招收 (“recruit”) and 学生 (“students”). While the first character of 招收 (i.e. 招) is chosen to combine with 生 to mean “students recruitment” in Mainland China, the preferred combination in Singapore is to use the second character of 招收 (i.e. 收) with 生.

Similarly, 节省 (“save”) and 水 (“water”) can result in either the compound words 节水, which is used more commonly in Mainland China, and 省水 in Singapore

(b) Common words, but different meanings:

<i>Words</i>	<i>Definition in Singapore</i>	<i>Definition in Mainland China</i>
财路 (cáilù)	Giro, a funds transfer banking system	Ways to make money
大字报 (dàzìbào)	Notices to demand repayment of debt	Big propaganda posters during the Cultural Revolution
同志 (tóngzhì)	Homosexuals	Comrades
劳改 (láogǎi)	“Corrective Work Order”(CWO), a law put in force in 1993 to punish litterbugs. Offenders are sentenced to cleaning up public facilities such as gardens or recreational parks	Forced labor as a form of punishment during the Cultural Revolution

Table 2.9: Examples of Words Defined Differently in Mainland China and Singapore

(c) Words from English influence:

The Chinese vocabulary used in Singapore is also rich in words that are created through direct transliteration of English words, e.g. 杯葛 (bēigé “boycott”), 固打 (gùdǎ “quota”), 罗厘 (luólǐ “lorry”), 积宝 (jībǎo “jackpot”), 德士 (deshì “taxi”), 史古打 (shǐgǔdǎ “scooter”), 固本 (gùběn “coupon”), 胡姬 (hújī “orchid”), etc. These words are usually not present in the vocabulary of Mainland China.

2.3.5 Transliteration Differences

The transliteration of non-Chinese proper names is often observed to vary amongst the Chinese speaking countries and regions. Here are some examples based two Chinese newspapers, namely *Lianhe Zaobao* of Singapore, and *People’s Daily* of Mainland China:

<i>Proper Names in English</i>	<i>Singapore</i>	<i>Mainland China</i>
AIDS	爱之病, 爱滋病	艾滋病
The Association of Southeast Asian Nations (ASEAN)	亚西安	东南亚国家联盟 (东盟)
Yeltsin, Boris	耶尔辛	叶利钦
Sihanouk, Norodom (King of Cambodia)	西哈诺	西哈努克
New Zealand	纽西兰	新西兰
European Union	欧洲联盟	欧洲共同体 (欧共体)
South China Sea	南中国海	南海
Summits, Summit Meetings	峰会	最高级会谈

Table 2.10: A Comparison of Transliterations of Non-Chinese Proper Names

2.4 Implications on Information Processing and Retrieval

As can be seen, the Chinese language is, in many ways, different from the Indo-English languages. Firstly, there are no spaces to separate words in a sentence; secondly, words are formed by one or more morphemes chosen from a repertoire of thousands of characters, and thirdly, the language possesses a much larger number of homophones and homographs, as compared to the Indo-European languages.

Two standards of character forms are used in today's Chinese writing, namely the traditional Chinese characters in Taiwan, Hong Kong and Macau, and the simplified Chinese characters in Mainland China and Singapore. Besides, political and sociological factors have been observed to affect the use of words amongst Chinese speaking communities, and the transliterations of non-Chinese proper nouns have not been standardized.

In view of processing Chinese information on computer systems, there are several issues to be addressed. Firstly, the number of keys in a standard western keyboard (e.g. QWERTY) is not adequate for the thousands of Chinese characters; and secondly, the standardization of an encoding system and input method for the Chinese characters may seem quite impossible, as there exist different systems of writing and character sets.

Additionally, the unclear definition of a word in the Chinese language, and the non-standardized transliteration methods complicate the process of indexing and information retrieval.

All in all, the complexity and inconsistency in the use of Chinese language is a challenge for Chinese text processing and information retrieval.

3 Chinese Information Processing

The processing of Chinese characters on computers involves some special issues in both character encoding schemes and character input. The purpose of this chapter is to give an overview of the existing encoding standards and inputting methods for Chinese characters, as well as to briefly describe the Chinese environment setup in a few operating systems, including *SuSE Linux Enterprise 8.0*, the operating system which was used for the experiments in this thesis.

3.1 Coded Character Set Standards and Encoding Methods

3.1.1 Introduction

Both coded character set standards and encoding methods are designed for information interchange and data communication on computer systems. While the coded character set standards define the repertoire of characters to be processed and used on computers, the process of encoding maps each of these characters to a numeric value, to create the ability to uniquely identify a character through its associated numeric value (Lunde (1999)).

For the fact that Chinese characters could be written in either its traditional or simplified forms (see Section 2.2.2), and that different Chinese character sets are standardized for common use in different countries and regions (see Section 2.3.1), the resulting coded character set standards for the Chinese language are also derived primarily at a national level, for example, the GB (abbreviation for 国家标准 or guójiā biāozhǔn, meaning “national standard”) character sets for the simplified Chinese characters developed by Mainland China, and the widely implemented Big-5 (大五碼 dàwǔmǎ) character set for the traditional Chinese characters used in Taiwan and Hong Kong. Similarly, there is no universally recognized encoding method for the Chinese characters, and many of the encoding methods used today are locale-specific.

The “Han Unification” action driven by the *Unicode Consortium* can be seen as the first effort to unify Chinese characters that are used in Mainland China, Taiwan, Hong Kong, Japan, Korean and Vietnam.

Today, both simplified and traditional characters could be displayed simultaneously by using the so-called unified encoding standards such as UTF-8 and GB 18030-2000, where both simplified and traditional Chinese characters are encoded. However, the conversion from simplified characters in the GB encoding to traditional characters in Big-5 encoding remains to be non-trivial, as a simplified character may map to multiple traditional equivalents (see Section 2.2.2).

3.1.2 Unicode, ISO 10646 and UTF Encodings

The Unicode Standard is a universal character set standard aimed at encoding all major languages, including Chinese. The first set of 20,902 unified Chinese characters, referred to as the “CJK Unified Ideographs”, was published concurrently in Unicode Version 1.0 and ISO 10646-1:1993 to compile Chinese characters that are commonly used in the written Chinese, Japanese and Korean languages. Subsequent extensions were made in 1998 and 2000 to include rarer and non-unifiable Chinese characters, namely the “CJK Unified Ideographs Extension A” in Unicode 3.0 – an addition of 6,582 Chinese characters from various industrial standards and historical literature, and the “CJK Unified Ideographs Extension B” in Unicode 3.1, that has 42,711 additional characters derived from major classical dictionaries and literary sources, as well as many national standards. To sum up, the Unicode Standard, at the time of this writing, has a repertoire of over 70,000 unified Chinese characters in, and is a superset of all existing Chinese national character sets, including GB 18030-2000 and Big-5.

The unification rules used to merge Chinese characters from the different national character sets are as follows (Unicode Consortium (2003): The Unicode Standard Version 4.0, pp. 300-301):

- R1. Source Separation Rule: If two ideographs are distinct in a primary source standard, then they are not unified.
- R2. Noncognate Rule: In general, if two ideographs are unrelated in historical derivation (noncognate characters), then they are not unified.
- R3. Any two ideographs that possess the same abstract shape are then unified provided that their unification is not disallowed by either the Source Separation Rule or the Noncognate Rule.

For example:

Characters	Reason
崖 ≠ 厓	Different number of components
峰 ≠ 峯	Same number of components placed in different relative position
擴 ≠ 擴	Same number and same relative position of components, corresponding components structure differently
區 ≠ 區	Characters treated differently in a source character set
祕 ≠ 秘	Characters with different radical in a component
爲 ≠ 為	Same abstract shape, different actual shape

Table 3.1: Ideographs Not Unified (Source: The Unicode Standard 4.0, p. 302)

Characters	Reason
周 ≈ 周	Different writing sequence
雪 ≈ 雪	Differences in overshoot at the stroke termination
酉 ≈ 酉	Differences in contact of strokes
鉅 ≈ 鉅	Differences in protrusion at the folded corner of strokes
璽 ≈ 璽	Differences in bent strokes
朱 ≈ 朱	Differences in stroke termination
父 ≈ 父	Differences in accent at the stroke initiation
八 ≈ 八	Difference in rooftop modification
說 ≈ 說	Difference in rotated strokes/dots

Table 3.2: Ideographs Unified (Source: The Unicode Standard 4.0, p. 302)

Unicode characters can be encoded in UTF-8, UTF-16 or UTF-32, where UTF stands for “Unicode Transformation Format” or “UCS Transformation Format” (as used in ISO 10646), and each UTF encoding uses a different code unit. For example, UTF-8, which is most commonly used on the Web, uses an eight-bit variable-length encoding, and UTF-16 and UTF-32 use 16 and 32 bits, respectively.

3.1.3 National Character Set and Encoding Standards

(a) Mainland China

GB 2312-80, which covers 6,763 commonly used simplified Chinese characters, was published in 1981 as the primary coded character set standard of Mainland China. In 1995, GBK or “Chinese Internal Code Specification” (国家标准扩展 guójiā biāozhǎn kuòzhǎn) was published as an extension to the existing GB standard. The character set was expanded to encode all 20,902 CJK unified ideographs that were assigned in Unicode 2.1⁷.

In order to provide more coded Chinese characters to meet the needs in Chinese information processing and interchange, as well as to ensure consistency in the encoding of Chinese characters used in both Chinese and other Minority Nationalities’ languages (e.g. Tibetan and Mongolian), the GB 18030-2000 characters set and encoding standard was published in March 2000. This new standard is a superset of GBK and GB 2312-80, and contains more than 27,000 Chinese characters which were defined in Unicode 3.0.

⁷ The CJK unified ideographs of Unicode 2.1 are identical to those of Unicode 1.0

All characters that are present in GB 2312-80 and GBK have exactly the same code assignment in GB 18030-2000, and this ensures compatibility across the different GB standards. However, the encoding in GB 18030-2000 is totally different from that of Unicode 3.0, although they share the same character repertoire.

The GB 18030-2000 adopts a one-byte, two-byte and four-byte encoding system, and provides over 1.5 million possible code points. Currently, 500,000 code points in GB 18030-2000 are still unassigned.

Mainland China has also mandated that any software application to be released for the Chinese market after September 2001 must support GB 18030-2000.

(b) Taiwan

CNS 11643-1992 is Taiwan's national character set and encoding standard, and contains 48,717 characters, including those for Japanese, Korean and simplified Chinese. However, this standard is seldom implemented. Instead, Big-5 (大五碼 dàwǔmǎ), a standard first released in 1984 by the "Institute for Information Industry of Taiwan" (台灣資訊工業策進會 táiwān zīxùngōngyè cèjìn huì), has become a *de facto* standard for Taiwan due to its long standing use on MacOS and Windows (Lunde (1999)). The original version of Big-5 (or Big5-1984) contained 13,053 traditional Chinese characters and 441 non-Chinese characters.

Over the years, several proprietary versions of Big-5 extensions were developed to replace Big5-1984, and it is not possible to single out one standard Big-5. The most widely used Big-5 extensions in Taiwan today include Big5-ETen (倚天版本 yǐtiān bǎnběn), Microsoft's CP950, and Big5-IBM. Official extensions of Big-5, namely Big5+ and Big5E, were introduced in 1997 and 1999, respectively. However, these official extensions are generally not well supported by the commercial and software industries, and are therefore rarely implemented in Taiwan. In 2003, a taskforce consisting of IT specialists and industrial representatives was formed by Taiwan's "Standards Board of the Ministry of Economics" (經濟部標準檢驗局 jīngjībù biāozhǔn jiǎnyànjú) to unify the different versions of Big-5. This unified version is referred to as Big5-2003.

(c) Hong Kong

Hong Kong also uses Big-5 for character encoding. However, the normal Big-5 character set does not contain all the special Chinese that are used in Hong Kong for person and place names, as well as characters unique to the Cantonese dialect. To solve this problem, the Hong Kong Government created the Big-5 extension "Government Chinese Character Set" (GCCS) in 1995. A revision, namely the "Hong Kong Supplementary Character Set" (HKSCS), was published in 1999 to contain 4,702 characters, and another 161 characters were added in 2001.

(d) Singapore

Singapore does not have its own character set and encoding standard. Both GB 2312-80 and GBK, though not as official standards, are widely used in the country. A move to the latest GB 18030-2000 standard has yet to be observed in Singapore. On the other hand, the use of ISO 10646 (i.e. Unicode) has been recommended by the *Information and Technology Standards Committee (ITSC)* of the Singapore's *Standards, Productivity and Innovation Board*, which is also a participating member of the "ISO/IEC Joint Technical Committee 1 on Information Technology".

3.2 Input Methods

There are three main types of keyboard input methods for Chinese characters:

- (a) by direct entry of character code values
- (b) by pronunciation (i.e. Pinyin and Zhuyin)
- (c) by structure of characters (e.g. by radicals, number of strokes, etc.)

Again, there is no single standard input method, and different people are most comfortable with different methods, as each method has its strengths and weaknesses.

3.2.1 Direct Entry of Character Code Values

This is an unambiguous but tedious input method, where the encoded values of the target characters are entered directly. For example, the character 字 may be entered as the row-cell number "55-44" for GB2312-80, or as the hexadecimal code "5B57" for Unicode, depending on the encoding system chosen for character code entry.

3.2.2 Zhuyin and Pinyin Input Methods

Both Zhuyin (or Bopmofo) and Pinyin input methods are based on the pronunciation of characters. These methods are more intuitive and are most frequently used to enter Chinese characters. Both methods involve the following steps:

1. Enter the Zhuyin symbol or Pinyin of a target character.
2. Select the target character from a candidate-list of characters that have the same pronunciation as the Zhuyin symbol or Pinyin entered

Note that while Pinyin can be entered by means of normal western keyboards, the entry of Zhuyin symbols require another set of keyboard array, such as the one shown in Figure 3.1.



Figure 3.1: Zhuyin Keyboard Array

To date, there are many different variations of sophisticated Chinese character input software that are based on the principal of pronunciation. These variations are not only capable of converting one character at a time, but compound words (i.e. a string of two or more characters), phrases, and even sentences.

Here are examples on Pinyin input for a single character 汉 and a compound word 汉字:

<i>Input Type</i>	<i>Target</i>	<i>Pinyin Input</i>	<i>Candidates in Selection List</i>
Single Character	汉	han	1 汉 2 韩 3 含 4 喊 5 邯 6 旱 7 翰 8 汗 9 寒 10 函 11 涵 12 罕 13 撼 14 焊 15 憇 16 瀚 17 酣 18 捍 19 憾 20 哈 21 颌 22 菡 23 蚶 24 邗 25 鼾 26 殢 27 烩 28 撤 29 阍
Compound Word	汉字	hanzi	1 汉字 2 汉子 3 汗渍 4 蚶子 5 憇子

Table 3.3: Pinyin Input Method for Single Characters and Compound Words

The Pinyin method mentioned so far is also referred to as “Full Pinyin” (全拼 quánpīn), where the Pinyin equivalent of Chinese characters are entered directly. There are two other types of Pinyin input methods which are called “Half Pinyin” (简拼 jiǎnpīn) and “Double Pinyin” (双拼 shuāngpīn). These two variations basically reduce the number of key strokes required to input a character, by using single- or double-letter codes to represent Pinyin consonants and vowels which are two or more letters long. For example, the Pinyin “zh”, “ch”, “ang”, “eng”, etc. are all reduced to single letter codes “a”, “u”, “h”(“g”), and “g”(“t”) in the “Half Pinyin” and “Double Pinyin” input methods. Table 3.4 illustrates the differences in using the three Pinyin input methods.

<i>Target</i>	<i>Full Pinyin</i>	<i>Half Pinyin</i>	<i>Double Pinyin</i>
啊	a	a	a
它	ta	ta	ta
沙	sha	ua	la
算	suan	su	sc
上	shang	uh	lg
双	shuang	uuh	lh

Table 3.4: Pinyin Input Examples

3.2.3 Input by Character Structure

Input methods by character structure employ techniques based on the radicals (部首 *bùshǒu*) of characters, or the strokes (笔画 *bǐhuà*) which formed the characters. The difference between radicals and strokes can be demonstrated using the character 字. This character has two radicals, 宀 and 子, but a total of 5 strokes, namely 丶, 丶, 冫, 了 and 一.

As compared to the direct encoding and Pinyin input methods, most input methods by character structure make use of less keys on a keyboard. For example, the “Five Stroke Method” (五笔划 *wúbǐhuà*), uses only the numerical keypad for character input. More specifically, it uses only the number keys 1 to 5, where “1” is assigned for horizontal strokes (一), “2” for vertical strokes (丨), “3” for diagonal downwards strokes from right to left (丿), “4” for short diagonal downwards strokes from left to right (丶), and “5” includes all other unassigned strokes such as 乙, 乚, ㇇, ㇈, ㇉, ㇊, etc. The input of a character is done by keying the numbers that correspond to the first four strokes and the last stroke of the character. For characters with four strokes or less, a 0 is entered after the last stroke. For example, the character 五 which is written in the sequence 一, 丨, 丿, and 丶, is entered as 1521, and the character 力 of stroke sequence ㇇ and ㇈, is entered as 530.

Similarly, the “Cangjie Input Method” (仓颉输入法 *cāngjié shūrùfǎ*), which allows combination input based on radicals, strokes or row-cell numbers of characters, uses only A to Y of the alphabet.

At present, there are different variations of input methods using character structure, and some allow even input using combination of pronunciation and character structure. Further examples are shown in Table 3.5.

<i>Input Methods by Character Structure</i>	<i>Input Methods by Pronunciation and Character Structure</i>
<ul style="list-style-type: none"> ▪ Five Stroke Character Form (五笔字形 wǔbǐ zìxíng) ▪ Head-Tail (首尾 shǒuwěi) ▪ Zheng-Code (郑码 zhèngmǎ) ▪ Dayi (大易 dàyì) ▪ Four Corners Code (四角码 sìjiǎomǎ) 	<ul style="list-style-type: none"> ▪ Tze-loi (子来 zǐlái) ▪ Renzhi Code (认知 rènzhī)

Table 3.5: Input Methods by (a) Character Structure, and (b) Pronunciation and Character Structure

3.2.4 Other Input Methods

Other means of inputting Chinese characters, though not commonly used, are through optical character recognition (OCR), voice recognition, and handwriting recognition software.

3.3 Chinese Environment Setup

The viewing and processing of Chinese texts on computers can be either done in a fully Chinese localized operating system, or in a “hybrid environment”, where Chinese support is enabled through the installation of special software and the use of a Chinese localized operating system is not required.

An overview of Chinese supporting features of a few operating systems, namely Windows®, MacOS, Unix and Linux, shall be covered in this section.

3.3.1 Microsoft® Windows®

Microsoft® Windows® operating systems⁸ are Unicode-based, and support three different Chinese localized versions, namely *Chinese Hong Kong*, *Chinese Simplified* and *Chinese Traditional*. Each version is bundled with Chinese fonts and a few input methods for the display and input of Chinese characters, and their individual features and differences are summarized in Table 3.6.

Though the default character set for simplified Chinese is GBK, GB 18030-2000 can be supported through the installation of the Microsoft’s “GB 18030 Support Package”.

⁸ Refers to Microsoft® Windows® 2000, Windows® NT, and Windows® XP (Home and Professional editions)

<i>Localized Version</i>	<i>Chinese Hong Kong</i>	<i>Chinese Simplified</i>	<i>Chinese Traditional</i>
<i>Locale Served</i>	Chinese (Hong Kong SAR) Chinese (Macau SAR)	Chinese (PRC) Chinese (Singapore)	Chinese (Taiwan)
<i>Supported Character Sets</i>	HKSCS GBK Big-5	GBK	Big-5
<i>Chinese Character Input Methods</i>	CangJie Cantonese ZhengMa MS Pinyin IME	NeiMa QuanPin ShuangPin ZhengMa MS Pinyin IME	Array Big5-Code CangJie DaYi Quick Unicode Zhuyin (phonetics) MS New Phonetic IME

Table 3.6: Chinese Localized Versions of Microsoft Windows® 2000, NT and XP

Besides using localized versions of Microsoft® Windows®, Chinese language support can be enabled in any localized versions of Windows® by installing the necessary files for East Asian languages, and by specifying one or more Chinese input languages (e.g. Chinese (Taiwan), Chinese (Hong Kong SAR), etc.) and the corresponding input methods through the “Add other languages” feature within the “Date, Time, Language and Regional Options” setting.

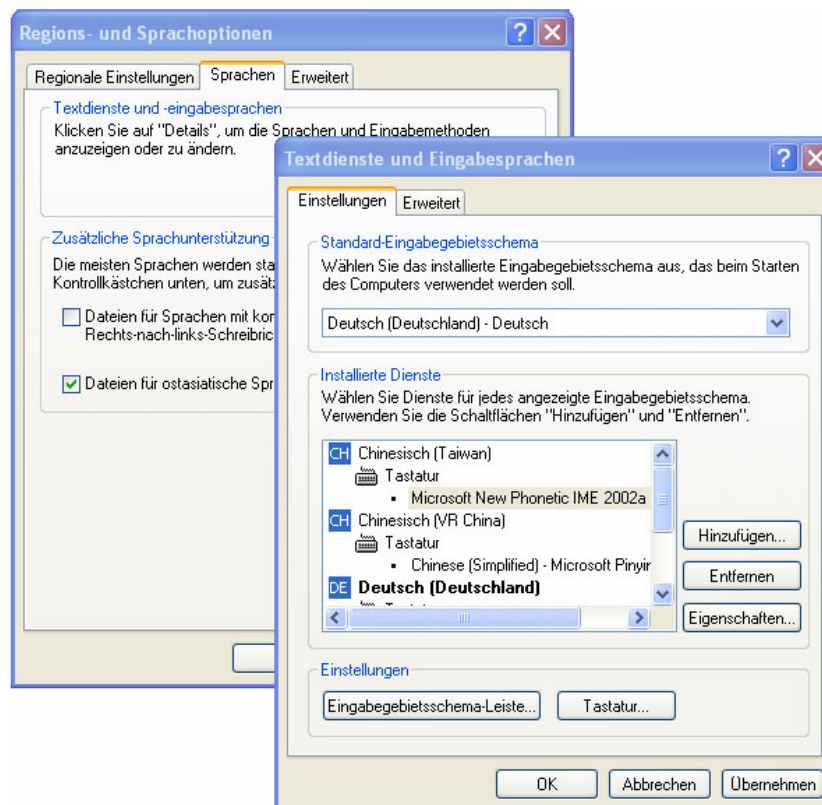


Figure 3.2: Enabling Chinese Support in Microsoft® Windows® Operating Systems

Once the settings are completed, Chinese can be selected as an input language from the “Language Bar”.



Figure 3.3: Selecting an Input Language from the “Language Bar” from a German Operating System

Additionally, a “Multilingual User Interface (MUI) Pack” has been recently introduced in the English version of *Windows® XP Professional*, to allow users to switch the language of the user interface, meaning names of menu options, choices in dialog boxes, and help system, to any of the 24 localized language versions of Windows.

3.3.2 MacOS

Like Microsoft®, Apple Macintosh computers are also Unicode-based, and can support simplified and traditional Chinese through either the fully localized versions of its operating system (MacOS), or through the installation of the “Chinese Language Kit” (CLK) software. The language setup and input methods in MacOS are very similar to those described in the previous section, and will therefore not be repeated here. A detailed account, however, is available at “Chinese Mac Frequently Asked Questions”⁹, a Website sponsored by the “Council on Asian Studies” of Yale University to support the use of Chinese language on Macintosh computers.

3.3.3 Unix and Linux

In both Unix and Linux operating systems, Chinese language support is enabled through the customization of the “locale”, i.e. the setting of locale specific environment variables such as `LC_TYPE` (character classification and case convention), `LC_NUMERIC` (non-monetary numeric formats), `LC_COLLATE` (collation order used for comparing and sorting), `LC_MONETARY` (monetary formats), `LC_MESSAGES` (language for systems messages), etc.

⁹ <http://www.yale.edu/chinesemac/index.html>

Instead of customizing the environment variables individually, the LANG environment variable, which has the syntax `language_territory.codeset`, can be used to perform a global change to all variables of a locale.

Take *SuSE Linux* for example, either the command (3.1) or (3.2) can be used to set the locale to handle simplified Chinese characters and the cultural aspects of Mainland China.

```
export LANG = zh_CN.GB18030 (3.1)
```

```
setenv LANG zh_CN.GB18030 (3.2)
```

In these commands, `zh` and `CN` represent “Chinese language” and “China”, respectively, and `GB18030` is the coded character set to be used. When all the locale categories are listed, it will show the following:

```
LANG=zh_CN.GB18030
LC_COLLATE=zh_CN.GB18030
LC_TYPE=zh_CN.GB18030
LC_MONETARY=zh_CN.GB18030
LC_NUMERIC=zh_CN.GB18030
LC_TIME=zh_CN.GB18030
LC_MESSAGES=zh_CN.GB18030
LC_ALL = (3.3)
```

The display and input of Chinese characters require the running of an “X Window” system (or some other Unix windowing environment) and an “X Input Method” (XIM) server. While most Unix and Linux systems are shipped with XIM servers to support CJK input, a couple of Chinese XIM servers are also available for free download from the WWW. These are, for example, the “Smart Common Input Method” (SCIM) platform that supports a “Smart Chinese Pinyin” input method, and *xcin* that supports both simplified and traditional Chinese, as well as an array of six input methods, namely “Double-Pinyin” (双音 shuāngyīn), “Cangjie” (仓颉 cāngjié), “Zhuyin” (注音 zhùyīn), “Five-Strokes-Method” (五笔 wǔbǐ), “Internal Codes” (内码 nèimǎ) and “Pinyin” (拼音 pīnyīn).

SuSE Linux Enterprise Server 8.0, which was used in the experiment of this thesis, had *xcin* as its default XIM server for all Chinese locales. This means, when a Chinese locale is set, for example the one shown in (3.3), *xcin* will be automatically started on system bootup and a so-called “OverTheSpot Window” (a small blue window) for Chinese input will be displayed on the desktop. “CTRL SPACE” can then be used to toggle the input mode between English and Chinese, and “CTRL-SHIFT” allows the selection of different Chinese input methods.

4 Indexing Methods for Chinese IR Systems

In most IR systems designed for the Indo-European languages, automatic indexing is done by extracting words or short phrases directly from the documents to be indexed. In comparison, indexing in Chinese IR systems require more sophisticated methods because word boundaries or words are not clearly defined in the Chinese language. The main approaches to indexing Chinese texts are classified as:

- (a) Character-based Indexing (or 1-gram Indexing)
- (b) N-gram Indexing
- (c) Word-based Indexing

As each of these approaches has its advantages as well as drawbacks, most of the recent works on Chinese IR have been based on hybrid indexing methods that combine two or more of these indexing approaches. For example, mixing character-based and bi-gram indexing could produce very good retrieval effectiveness, but has the disadvantage of entailing high storage costs; and combining character- and word-based indexing has been shown to give a good balance between retrieval effectiveness and efficiency (Kwok (1999); Luk/Kwok (2002)).

4.1 Character-based Indexing (1-gram Indexing)

This is the simplest Chinese indexing method, whereby neither a stoplist, dictionary of pre-defined terms nor linguistic analysis is required, and every character is considered an index token (Kwok (1999)).

This type of indexing has demonstrated to produce retrieval results with high recall, but low precision (Tong et al. (1996); Kwok/Grunfeld (1996)). As illustrated by Kwok (1999), a string such as 中共核电站之营运情况¹⁰ (“the operating conditions of the Chinese nuclear power plants”), when represented by the single-character tokens, namely 中, 共, 核, 电, 站, 之, 营, 运, 情, 况, can result in high imprecision when matched with queries, as each of the single-character tokens may contain additional meanings that are different from the original context of the phrase. For example, the token 中 can now mean “center”, when not joined with 共; 站 can also mean the verb “stand” or a “stopping place” like 巴士站 (“bus-stop”); and 核 can be a word by itself meaning “pit of a fruit”, etc.

¹⁰ This is a topic statement taken from TREC-5. The individual words 中共, 核电站, 营运, and 情况 mean “PRC” (expressed as an abbreviation for “Peoples’ Republic of China”), “nuclear power plant”, “operating”, and “conditions”, respectively, and 之 is a functional word.

On the other hand, the token 电 (meaning “electricity”) can be matched to thematically related words such as 电池 (“battery”), 闪电 (“lightning”), 电灯 (“electric lamp”), 电动 (“motor-driven”), etc. In effect, this is similar to “stemming” in English IR, where words belonging to the same root (e.g. computers, computerization) are matched to a query. Hence character-based indexing has the advantage of producing high recall in retrieval (i.e. having the ability to retrieve most of the relevant documents that are present in a given collection).

4.2 N-gram Indexing

N-gram indexing, like character-based indexing, does not require the use of a dictionary nor linguistic knowledge. Given a text, its index tokens are all possible consecutive overlapping n characters that are found in this text.

As most words in the modern Chinese language are disyllabic (see Section 2.3.2), “bi-gram indexing” is the most often applied n-gram indexing method, whereby all consecutive overlapping 2-character pairs form the tokens. For example, the same string used in Section 4.1 is now represented by the bi-grams 中共, 共核, 核电, 电站, 站之, 之营, 营运, 运情, and 情况.

Five of these bi-grams are real word tokens, namely 中共 (“PRC”), 核电 (“nuclear power”), 电站 (“electric station”), 营运 (“operating”), and 情况 (“conditions”). As compared to the single-character tokens in character-based indexing, these real word bi-grams are more specific in meaning, and can therefore be matched more precisely to queries.

On the other hand, the proportion of meaningless character-pairs (共核, 站之, 之营 and 运情) is also quite high – about 45% in this example.

One obvious drawback of bi-gram indexing is the inability to index single-character words. Take for example, 铁 (“metal”), which is a single-character word, cannot be indexed as a single token in the string 铁的产量 (“amount of iron production”) (Kwok (1999)).

Additionally, since n-1 tokens are generated for every string of n characters, bi-gram indexing can result in an enormous index for large collections, and will therefore take up a considerable amount of storage space.

4.3 Word-based Indexing

Word-based indexing involves the generation of real word tokens from segmented texts (i.e. texts with word boundaries). Usually, stopword lists (or stoplists) are also used to exclude functional words from the segmented texts, and linguistic and heuristic rules may be applied to expand or normalize the words of the segmented texts.

Most techniques used for automatic Chinese word segmentation can be broadly classified as either dictionary-based or statistically-based (Nie/Brisebois/Ren (1996)). Typically, dictionary-based word segmentation is based on the use of a comprehensive pre-compiled lexicon consisting of commonly used words, proper names and idioms. Given an input string, it is scanned from left to right to identify the longest possible word match to the lexicon. This matching procedure is referred to as the “greedy match” or “maximum match” algorithm. The statistical approaches, on the other hand, identify word boundaries or words through the use of statistical information such as word and character co-occurrence frequencies (Dai/Khoo/Loh (1999)), or complex statistical models (e.g. first-order Markov models).

In general, word-based indexing has the advantages of supporting stopwords, and generating smaller indices (as compared to bi-gram indexing) of real word tokens.

5 Retrieval Performance Evaluation

The retrieval effectiveness of an information retrieval (IR) system is commonly measured in terms of *precision* (accuracy of a search) and *recall* (comprehensiveness of a search).

This chapter shall first provide an overview of some commonly used IR evaluation measures, namely the recall-precision curve, the average precision and R-precision values, and the precision histogram. Following which, the standard IR performance evaluation method used in *Text Retrieval Conference* (TREC) shall be explained, and some issues involved in this method of evaluation will also be discussed.

5.1 Recall and Precision

Conceptually, recall is a measurement on the ability of an IR System to find all of the relevant items present in a collection. Precision, on the other hand, assesses the relevance of the retrieved items.

5.1.1 Definitions

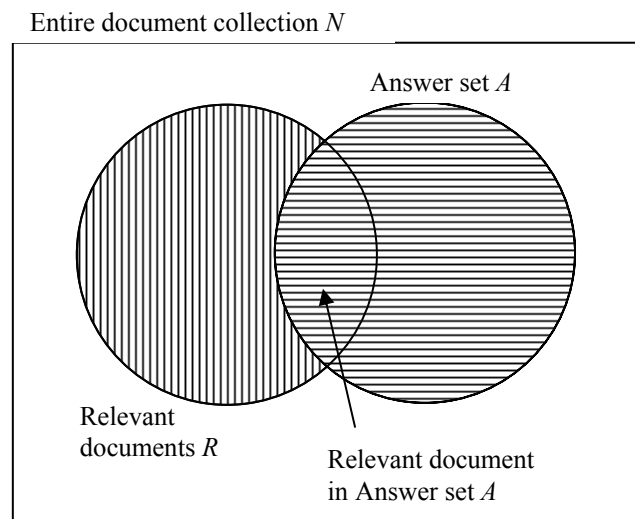


Figure 5.1: Recall and Precision for a Given Example Information Request (adapted from Baeza-Yates/Ribeiro-Neto (1999), p. 75).

Consider a query that is searched against a document collection N , and suppose Figure 5.1 represents the distribution of the relevant documents corresponding to this query (R), and the documents that are present in answer set (A), then the recall and precision measures are defined as follows:

- Recall is the fraction of the relevant documents which has been retrieved i.e.,

$$recall = \frac{|R \cap A|}{|A|} \quad (5.1)$$

- Precision is the fraction of the retrieved documents which is relevant i.e.,

$$precision = \frac{|R \cap A|}{|R|} \quad (5.2)$$

Ideally, both recall and precision should equal to one, meaning that all documents in the answer set A are relevant. However, empirical studies of retrieval performance have shown a tendency for precision to decline as recall increases (van Rijsbergen (1979)). The recall-precision curve in Figure 5.2 illustrates this observation.

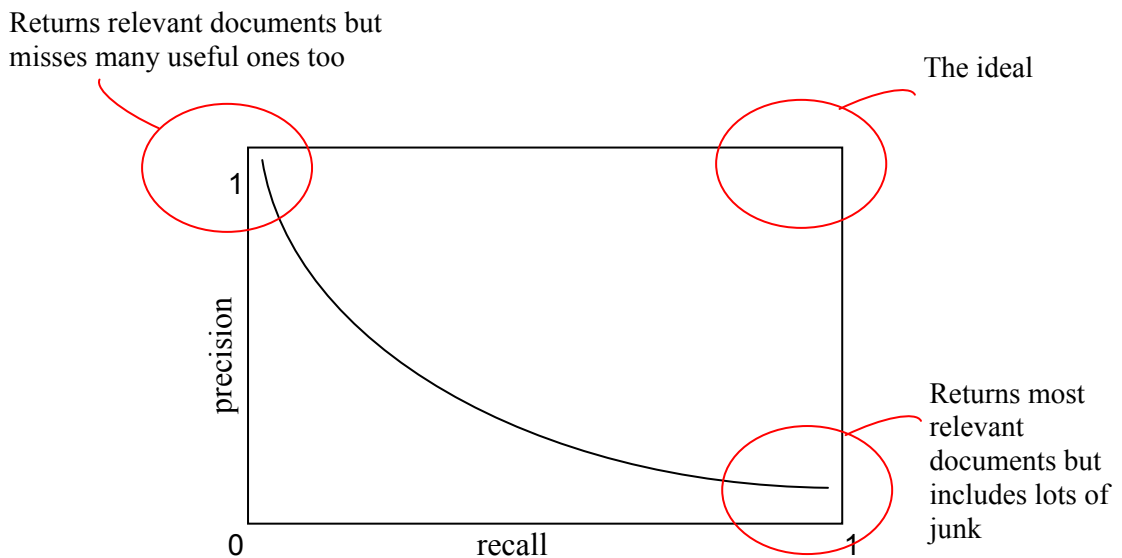


Figure 5.2: Trade-off Between Recall and Precision

5.1.2 Recall-Precision Curve and Precision Interpolation

The most commonly used method to evaluate the effectiveness of an IR system is to plot a recall-precision curve on 11 standard recall-points (0.1, 0.2, 0.3,..., 1.0). The precision value corresponding to each of these recall-points is interpolated as follows:

$$P(r_j) = \max_{r_j \leq r \leq r_{j+1}} P(r) \quad (5.1)$$

where

$$r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$$

and $r_0 = 0.0, r_1=0.1, r_2=0.2, \dots, r_{10}=1.0$

Example:

Assume a ranked list of documents A_q is retrieved for a query q , and let this ranked list and all relevant documents corresponding to this query R_q be represented as follows:

$$A_q = \{d_{123}, d_5, d_{56}, d_{39}, d_6, d_9, d_{44}, \dots, d_3, d_{13}\} \quad (5.2)$$

$$|A_q| = 14$$

$$\text{and } R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{123}\} \quad (5.3)$$

$$|R_q| = 6$$

The interpolation of precision values for the 11 standard recall points is then computed as follows:

Step 1:

Identify all relevant docs within the ranked list of retrieved documents

Step 2:

Calculate the recall and precision values for all relevant docs found

Step 3:

Interpolate the precision values using equation (5.1) for all 11 recall points

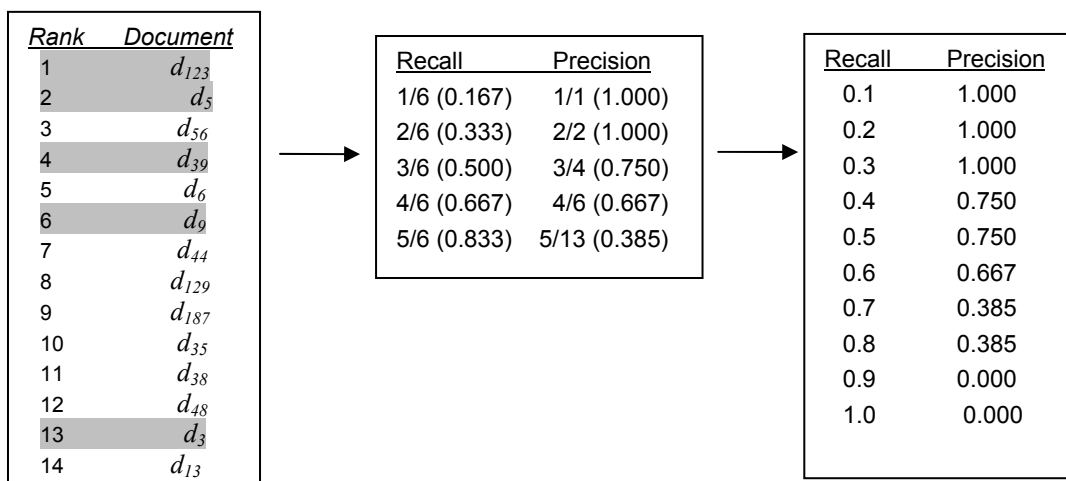


Figure 5.3: Computation of Recall-Precision Points (non-interpolated)

Note that in this example, not all relevant documents in the collection have been retrieved. As a result, recall levels higher than 0.833 are not found in the non-interpolated results, and the interpolated precision values for recall levels 0.9 and 1.0 are therefore estimated to be zero.

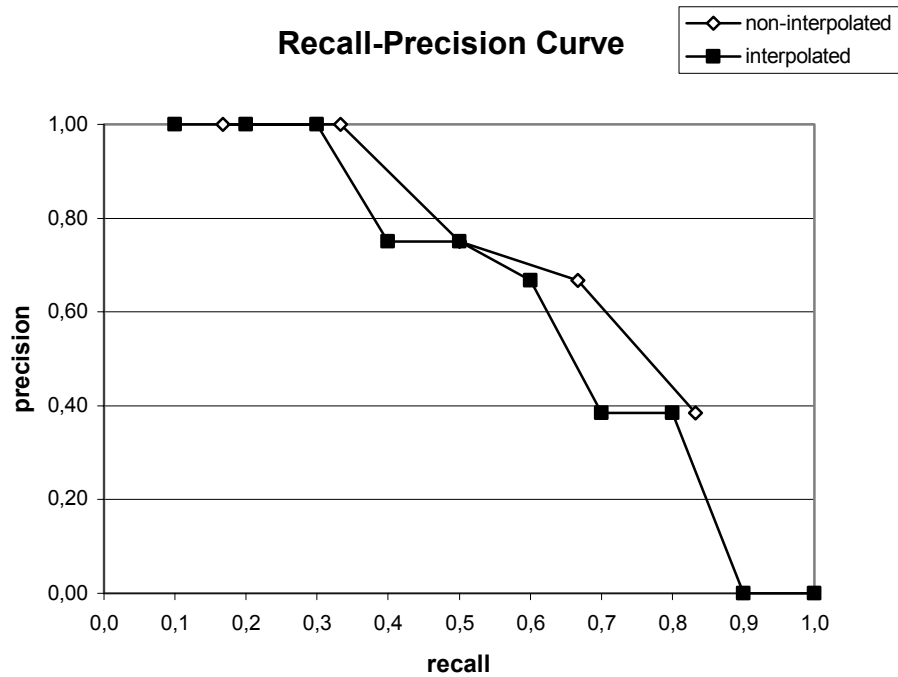


Figure 5.4: Recall-Precision Curve (interpolated and non-interpolated)

Typically, recall-precision curves are plotted from retrieval results based on several queries. In this case, the interpolated precision values are averaged as follows:

$$\bar{P}(r_j) = \sum_{i=1}^{N_q} \frac{P_i(r_j)}{N_q} \quad (5.5)$$

where $\bar{P}(r_j)$ is the average precision at the recall level r_j , N_q is the number of queries used, and $P_i(r_j)$ is the precision at recall r_j for the i -th query.

5.1.3 Average Precision and R-Precision

Average Precision and *R-Precision* are two commonly used single-value evaluation measures for IR systems. Making use of the same document ranked list shown in Figure 5.3, these values can be defined as follows:

<i>Single Value Measures</i>	Average Precision	R-Precision
<i>Definition</i>	Mean of precision values after each new relevant document	Ratio of relevant documents among the first $ R_q $ -th documents
<i>Example based on Figure 5.3</i>	$1+1+0.75+0.667+0.38/5$ $= 0.76$	$4/6$ $= 0.667$
<i>Interpretation</i>	A high value indicates good relevance ranking	A high R-Precision indicates high recall and good relevance ranking

Table 5.1: A Comparison of Average Precision and R-precision

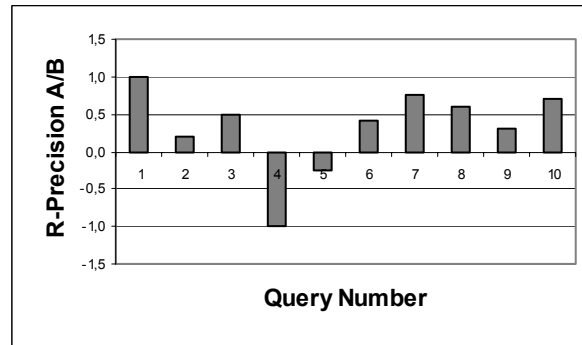
Both average precision and R-precision values are usually computed for individual queries to observe the behavior of an algorithm for each query. However, averages over all queries are also sometimes used to summarize the overall performance of an IR system.

5.1.4 Precision Histogram

The precision histogram, as shown in Figure 5.5, is a bar chart that compares the retrieval performance of two retrieval algorithms. The vertical-axis values, namely $(RP_{A/B}(i))$, are computed as follows:

$$RP_{A/B}(i) = RP_A(i) - RP_B(i) \quad (5.6)$$

where $RP_A(i)$ and $RP_B(i)$ are the R-precision values for the i -th query of algorithms A and B, respectively.



Observation:

Algorithm A performs better in queries 1, 2, 3, 6, 7, 8, 9, and 10, where $RP_{A/B}(i) > 0$,

but worse in queries 4 and 5, where

$RP_{A/B}(i) < 0$

Figure 5.5: A Precision Histogram for Ten Hypothetical Queries
(Baeza-Yates/Ribeiro-Neto (1999), p. 81)

5.2 The Text Retrieval Conference (TREC)

The *Text Retrieval Conference* (TREC) was started in 1992 to “support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies”¹¹

In each TREC, participants are provided with standard test collections and topic statements (questions) to perform retrieval tasks defined for the conference workshops. All retrieval results submitted by the participants are then evaluated, and the effectiveness of different retrieval techniques is compared.

5.2.1 Tasks and Tracks

Two main tasks are carried out in each TREC workshop:

- (a) *Routing task*: the performance evaluation of systems that use standing queries to search new streams of documents (i.e. similar to IR required by news clipping and library profiling systems)
- (b) *Ad-hoc task*: The performance evaluation of systems that search a static set of documents using different topics (i.e. similar to IR for library reference-type questions). All topic statements, like the sample topic statement shown in Figure 5.6., are expressed in both short (title and description) and detailed (narrative) versions. Query construction, on the other hand, can be either done *automatic* or *manual*. While the former involves completely automatic query construction from the topic statements, the latter does not impose any constraints on query construction. Revision of queries after examining the retrieved documents is allowed in the *manual* method.

¹¹ National Institute of Standards and Technology (NIST): Text Retrieval Conference Homepage. <http://trec.nist.gov> (Date of access: 22.01.2004)

```

<top>
<num> Number: CH1
<E-title> U.S. to separate the most-favored-nation status from human
rights issue in China.
<C-title> 美国决定将中国大陆的人权状况与其是否给予中共最惠国待遇分离.
<E-desc> Description:
most-favored nation status, human rights in China, economic sanctions,
separate, untie
<E-narr> Narrative:
A relevant document should describe why the U.S. separates most-
favored nation status from
human rights. A relevant document should also mention why China
opposes U.S. attempts
to tie human rights to most-favored-nation status.
<C-desc> Description:
最惠国待遇, 中国, 人权, 经济制裁, 分离, 脱钩
<C-narr> Narrative:
相关文件必须提到美国为何将最惠国待遇与人权分离; 相关文件也必须提到中共为什么反对美国
将人权与最惠国待遇相提并论.
</top>

```

Figure 5.6: Sample Topic Statement for *Ad-hoc* Retrieval in the Chinese language.

Besides the routing and ad-hoc tasks, secondary tasks (*tracks*) have also been introduced since TREC-4 to allow more specific comparisons. For example, the *Question Answering Track*, *Interactive Track*, *Cross-Language Track*, and *Multilingual Track*, etc, is each focused on a particular aspect of text retrieval.

5.2.2 Relevance Judgment

The *pooling* method is used in TREC to determine the relevance of documents. Given a topic statement, the top 100 ranked documents from each run¹² are first collected, and any duplicates in this pool are then removed. All remaining documents are then evaluated by human judges for their relevance to the topic. “Only binary judgments (‘relevant’ or ‘not relevant’) are made, and a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document)”¹³.

Two assumptions have been made in this technique of assessment:

- (a) most of the relevant documents in the entire collection are collected in the pool consisting of top 100 ranked documents from each run
- (b) documents which are not assessed are considered not relevant

¹² A run refers to a retrieval trial submitted by a participant for evaluation. Each run contains a set of 1000 ranked documents for each topic, and a participant may submit more than one run.

¹³ **National Institute of Standards and Technology (NIST): Data – Non-English Relevance Judgments.** http://trec.nist.gov/data/reljudge_noneng.html (Date of access: 22.01.2004)

5.2.3 Evaluation Measures

The evaluation instrument used by TREC is based on relevance, with recall and precision as the primary measure of effectiveness. The four evaluation measures used are:

- (a) *Summary Statistic* – a summary table that comprises the following statistics: (i) total number of topics; (ii) number of documents retrieved over all topics; (iii) total number of relevant documents retrieved; and (iv) total number of relevant documents.
- (b) *Recall-Precision Averages* – consists of: (i) recall-precision curve plotted on 11 standard recall-points; and (ii) mean of the non-interpolated average precisions over all topics.
- (c) *Document Level Averages* – includes: (i) average *R-precision* over all topics; and (ii) average precision over all topics at 9 document cut-off values, namely 5, 10, 15, 20, 30, 100, 200, 500 and 1000.
- (d) *Average Precision Histogram* – a precision histogram that measures the precision of a run on each topic against the median precision of all corresponding runs on that topic.

5.2.4 Chinese Document Retrieval in TREC-5 and TREC-6

Ad-hoc retrieval in the Chinese language (or Chinese Track) was first introduced in TREC-5 as a retrieval task in the *Multilingual Track*. 28 topic statements were provided to the participating groups to search against a Chinese newspaper corpus of 164,788 documents drawn from the *People's Daily* newspaper and the *Xinhua News Agency*. These documents are non-segmented simplified Chinese texts which are encoded in the GB2312-80 encoding scheme. The same task was repeated in TREC-6 using another 26 new topics. Table 5.2 summarizes the the key approaches adopted by the individual participants in TREC-5 and TREC-6 to perform the retrieval tasks.

Participants	TREC-5	TREC-6
City University	<ul style="list-style-type: none"> ▪ single character indexing ▪ dictionary based word segmentation based on a greedy algorithm and a 70,000-entry dictionary ▪ phrase weighting 	
Claritech Corporation	<ul style="list-style-type: none"> ▪ words/phrase indexing using morphological rules + 100,000-entry Chinese lexicon ▪ single character indexing ▪ overlapping character bi-grams 	<ul style="list-style-type: none"> ▪ bi-grams ▪ automatic feedback
Cornell University	<ul style="list-style-type: none"> ▪ single character indexing ▪ bi-grams 	
George Mason University	<ul style="list-style-type: none"> ▪ single character indexing ▪ term expansion 	
Information Technology Institute	<ul style="list-style-type: none"> ▪ single character indexing ▪ phrase retrieval performed by first retrieving the position lists of the constituent terms, followed by proximity check 	<ul style="list-style-type: none"> ▪ novel matching algorithm for character based retrieval using positional information ▪ expansion terms selected from 3-grams
Institute of Systems Science	<ul style="list-style-type: none"> ▪ bi-grams ▪ dictionary based segmentation + greedy (longest match) algorithm ▪ 2-4 grams 	
Royal Melbourne Institute of Technology	<ul style="list-style-type: none"> ▪ combination of single character, bi-grams and words found from dictionary methods 	<ul style="list-style-type: none"> ▪ combination of single character, bi-grams and words found from dictionary methods ▪ term expansion
Queens College, CUNY	<ul style="list-style-type: none"> ▪ combination of dictionary and statistical techniques to detect 2, 3 and occasionally 4 characters words ▪ combination of single characters and words 	<ul style="list-style-type: none"> ▪ combination of bi-grams and short words indexing using a 43K lexicon
Swiss Federal Institute of Technology (FIT)	<ul style="list-style-type: none"> ▪ single character indexing ▪ bi-grams ▪ term expansion 	<ul style="list-style-type: none"> ▪ bi-grams ▪ elimination of stopwords and identification of word boundaries through manually generated stop list of almost 1000 Chinese words
University of California, Berkeley	<ul style="list-style-type: none"> ▪ 140,000-entry dictionary to automatically segment text and queries based on longest match plus overlap match ▪ manual query modification by adding new words, changing term weights, and adding negative terms 	<ul style="list-style-type: none"> ▪ dictionary of 150,000 words to automatically segment text ▪ Local Context Analysis approach for term expansion
University of Massachusetts	<ul style="list-style-type: none"> ▪ Hidden Markov Model to segment text ▪ Local Context Analysis approach for term expansion 	
University of Montreal		<ul style="list-style-type: none"> ▪ sophisticated morphological analysis as word identification algorithm ▪ bi-grams
University of Waterloo		<ul style="list-style-type: none"> ▪ individual character indexing augmented by phrases based on adjacent characters using <i>Multitext</i> approach

Table 5.2: Indexing and Retrieval Approaches Used in the Chinese Track of TREC-5 and TREC-6

5.3 Limitations of TREC Evaluation

5.3.1 Relevance Assessments and Missed Documents

Both precision and recall measures assume that relevance is a meaningful concept and that relevance assessments possess the requisite stability on which valid performance measures can be constructed (Harter (1996)). This assumption has often been criticized, as relevance is a subjective notion, and experimental studies have shown that relevance judgments are affected by various factors, such as the characteristics of judges, requests, documents, information systems, judgment conditions and the choice of scale (Schamber (1994)). Moreover, relevance assessments employed in several experiments and test procedures have been reported to vary (Harter (1996)), and this further questions the reliability of the precision and recall measurements.

The binary relevance assessment used in TREC is also criticized for allowing a very low threshold for accepting a document as relevant, which does not allow the differentiation between systems that are capable of retrieving highly relevant documents, and those with only marginally relevant documents (Kekäläinen/Järvelin (2002)).

A further study done on the TREC-1 results suggested the possibility that many relevant documents were missed in the process of relevance assessment (Harter (1996)). This implies that recall cannot be estimated precisely, and the accuracy of the evaluation is affected negatively.

On the other hand, it has been proven that the comparative evaluation of retrieval performances is stable despite substantial differences in relevance judgment (Voorhees (1998)). High correlations were found among the rankings of systems produced using different relevance judgment sets. Hence, this justifies the use of recall and precision measures for cross-system comparisons in TREC.

5.3.2 Pooling and Averaging

The TREC evaluation employs the technique of pooling in relevance judgment, and the technique of averaging in the analysis of results (e.g. recall-precision averages, document level averages, etc). This, as observed by Harter (1996), erases the differences in how a system performs for different kinds of users, queries and their associated relevance judgment. Thus, the pooling and averaging approaches provide only a partial picture of the performance of an IR system, and it is necessary to look at the results of individual queries with individual systems (Tague-Sutcliffe (1996)).

6 Oracle® Text and Chinese IR

Oracle® Text is a text search tool that uses the standard *structured query language* (SQL) to index, search, and analyze text and documents stored in an *Oracle9i* or *Oracle10g* database.¹⁴

There are four index types offered in *Oracle® Text*:

- (a) *context* – a standard index type for traditional full-text retrieval over documents and Web pages
- (b) *ctxcat* – a catalogue index type designed for catalogue or inventory type information such as that of an online bookstore
- (c) *ctxrule* – a classification index type for building classification or routing applications
- (d) *ctxxpath* – a XPath index for improving performance on XPath searches on XML documents

For the purpose of this thesis, only the *context* index will be dealt with. To be covered in this chapter, are firstly, the database and *National Language Support* (NLS) settings required for Chinese indexing; secondly, the use of `CREATE INDEX` and index preferences to create a *context* index (i.e. full-text index) for Chinese documents; and lastly, some examples on the use of `CONTAINS` operator to search Chinese texts.

6.1 Database Character Sets and NLS Setting

The first step to indexing and searching Chinese texts in an *Oracle* database, is to choose a suitable database character set, and to set a suitable NLS environment for the client-database interaction.

Table 6.1 provides a list of *Oracle*-supported character sets for simplified and traditional Chinese. The choice of character set depends on the writing systems of the Chinese documents to be stored in the database.

¹⁴ **Oracle Corporation:** Official Page of Oracle Text in Oracle Technology Network. URL: <http://technet.oracle.com/products/text>. (Date of Access: 28.01.04)

<i>Writing system</i>	<i>Character Set</i>	<i>Description</i>
Simplified Chinese	ZHS32CGB18030	GB18030-2000, the mandatory standard in mainland China
	ZHS16CGB231280	GB2312-80, the most commonly used standard for simplified Chinese
	ZHS16GBK	GBK, an extension of the GB2312-80
	UTF8	8-bit encoding of Unicode, a universal character set
Traditional Chinese	ZHT32EUC	Extended Unix Code (EUC) 32-bit traditional Chinese
	ZHT16BIG5	BIG-5, the most commonly used standard in Taiwan and Hong Kong
	ZHT32TRIS	TRIS 32-bit traditional Chinese
	ZHT16MSWIN950	MS Windows Code Page 950 for traditional Chinese
	ZHT16HKSCS	HKSCS, an extension of Big-5 that is used in Hong Kong
	UTF8	8-bit encoding of Unicode, a universal character set

Table 6.1: Oracle-supported Character Sets for Chinese Characters

As mentioned in Chapter 3, different character sets support different character repertoires. So, it is important to choose a database character set that covers all the desired language needs. For example, the character set ZHS16CGB231280 supports only simplified Chinese characters, but not traditional Chinese characters. Hence, when ZHS16CGB231280 is chosen as the database character set, data will not be stored and retrieved properly for a client locale that uses traditional Chinese characters (e.g. ZHT16BIG5). A more flexible approach is to actually use UTF8 as the database character set, since Unicode 4.0, is at present, a superset of all existing national character sets for simplified and traditional Chinese. Figure 6.1 shows the NLS architecture for a UTF-8 based server that is connected to clients using different Chinese locales.

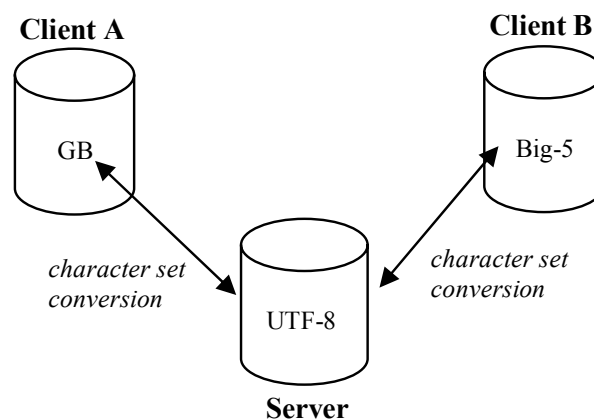


Figure 6.1: NLS Architecture for Storing, Processing and Retrieval of Data in Simplified and Traditional Chinese

The locale of a client is determined by the NLS parameters, which include the client preferences for time, date, calendar, numeric, monetary, collation (sorting), language, territory, and character set. The simplest way to specify locale behavior is to use the `NLS_LANG` parameter. This parameter, which has the syntax expressed in (6.1), sets the language, territory, and character set to be used for the server session and client application.

```
language_territory.charset                                (6.1)
(Default Value: AMERICAN_AMERICA.US7ASCII)
```

Here are some examples of `NLS_LANG` settings for different Chinese locales:

<i>Locale</i>	<i>NLS_LANG</i>
Mainland China	Simplified Chinese_China.ZHS32GB18030
Taiwan	Traditional Chinese_Taiwan.ZHT16BIG5
Hong Kong	Traditional Chinese_Hong Kong.ZHT16HKSCS
Singapore	Simplified Chinese_Singapore.ZHS16CGB231280

Table 6.2: `NLS_LANG` Settings for Different Chinese Locales

6.2 Indexing Chinese Text

6.2.1 Context Index

Consider a table *tab* that contains two text columns, *id* and *doc*. A *context* index can be created for the *doc* column as follows:

```
CREATE INDEX tab_index ON tab(doc)
  INDEXTYPE IS CTXSYS.CONTEXT;                                (6.2)
```

Optionally, index preferences may be included in the SQL command. For example:

```
CREATE INDEX tab_index ON tab(doc)
  INDEXTYPE IS CTXSYS.CONTEXT
  PARAMETERS('DATASTORE my_datastore LEXER my_lexer');      (6.3)
```

In command (6.3), the preferences in the parameter string, namely *datastore* and *lexer*, specify how the text to be indexed is stored and which language to use for indexing. Table 6.1 provides a list of all seven classes of index preferences which are supported by Oracle® Text. The rules and syntax (`CTX_DDL`) pertaining to the specification of preferences are, however, beyond the scope of this work.

Preference Class	Function
<i>Datastore</i>	Specifies how are the documents stored, e.g. stored internally in a text column, stored in a text table in more than one column, stored externally in operating system files, etc.
<i>Filter</i>	Specifies how the documents can be filtered for indexing, e.g. filters for formatted documents, plain text, <i>XML</i> , etc.
<i>Lexer</i>	Specifies the language of the documents being indexed, e.g. <i>Basic_lexer</i> (for English, German, French, etc.), <i>Chinese_lexer</i> and <i>Chinese_Vgram_lexer</i> for Chinese, etc.
<i>Wordlist</i>	Specifies how stem and fuzzy queries should be expanded.
<i>Storage</i>	Specifies how the index tables should be stored.
<i>Stoplist</i>	Specifies the words and themes not to be indexed.
<i>Section Group</i>	Specifies whether querying within sections is enabled, and how are the document sections defined.

Table 6.3: Preference Classes (adapted from *Oracle Corp.* (2003))

6.2.2 Lexer, Wordlist and Stoplist

When a *context* index is created on Chinese documents, it is important to ensure that the preferences for *lexer*, *wordlist*, and *stoplist* are specified correctly. Otherwise, the Chinese characters will not be indexed or searched properly.

(a) Chinese Vgram Lexer vs. Chinese Lexer

Oracle® Text supports two *lexers* for the indexing of Chinese texts, namely *Chinese_Vgram_Lexer* and *Chinese_Lexer*.

The *Chinese_Vgram_Lexer* uses overlapping bi-grams to break a character string into series of tokens, and every character is effectively an indexing point. The *Chinese_Lexer*, on the other hand, uses a hybrid algorithm that combines dictionary-based word indexing and bi-gram indexing. It first uses a "greedy algorithm" to find the longest match of words from a Chinese lexicon, and when no match can be made from the lexicon, a bi-gram algorithm is used. Figure 6.2 provides a flow chart for the hybrid algorithm adopted by the *Chinese_Lexer*.

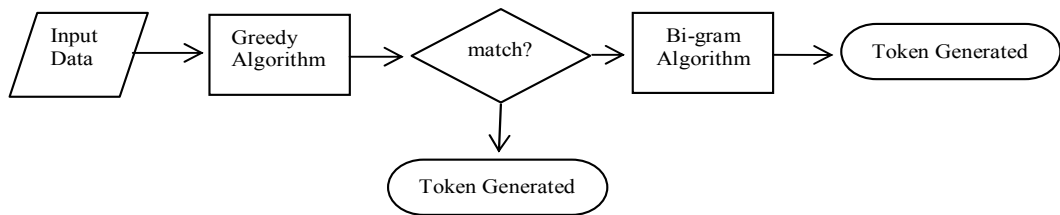


Figure 6.2: Hybrid Algorithm of *Chinese_Lexer* (Chen (2001))

In general, the *Chinese_Lexer* has the following advantages over the *Chinese_Vgram_Lexer*:

- generates a smaller index
- better query response time
- generates real word tokens, including country names and person names
- supports stopwords

Example:

Consider a character string 中文字符集的选择 (“selection of Chinese character set”). The tokens generated for this string using different lexer preferences are as follow:

<i>Lexer</i>	<i>Tokens</i>	<i>Remarks</i>
Chinese_Vgram Lexer	(1) 中文, (2) 文字, (3) 字符, (4) 符集, (5) 集的, (6) 的选, (7) 选择, (8) 择	Only tokens (1), (2), (3) and (7) are real word tokens
Chinese_Lexer (without stopwords)	(1) 中文, (2) 文字, (3) 字符, (4) 的, (5) 选择, (6) 集的	The real word tokens (1) – (5) are generated through the greedy algorithm, and token (6) is generated through bi-gram algorithm
Chinese_Lexer (stopword = 的)	(1) 中文, (2) 文字, (3) 字符, (4) 选择, (5) 集的	The stopwords 的 is not indexed

(b) Wordlist (stemmer and fuzzy match)

The wordlist preference for the Chinese language is set as follows:

```

begin
  CTX_DDL.create_preference('DEFAULT_WORDLIST','BASIC_WORDLIST');
  CTX_DDL.set_attribute('DEFAULT_WORDLIST','STEMMER','NULL');
  CTX_DDL.set_attribute('DEFAULT_WORDLIST','FUZZY_MATCH',
    'CHINESE_VGRAM');
end;

```

(6.4)

As the Chinese language is ideograph-based, no stemming is required in queries, and the *stemmer* attribute is therefore set to NULL. The *fuzzy_match* attribute value is a dummy attribute value that prevents the English and Japanese fuzzy matching routines from being used on Chinese text (Oracle Corp. (2003)).

(c) Stoplist (simplified and traditional Chinese)

Oracle® Text provides default stoplists for both simplified and traditional Chinese (see Appendix E). Each list contains 76 words which were selected from the most frequently used words in the lexicon applied in the greedy algorithm (Chen (2001)). Most of the stopwords are functional words and pronouns such as:

可能	maybe
以及	and
我们(我們)	we
他们(他們)	they
所有	all
因为(因為)	because
但是	but, however
必须(必須)	must

The stoplists for simplified Chinese and traditional Chinese are not identical. Some words are present in one list but not in the other. For example, the words 繼續 (continue), 加強 (strengthen or reinforce), and 尤其 (especially) are stopwords found in the traditional Chinese stoplist only. Similarly, the stoplist for simplified Chinese contains words such as 如何 (how), 获得了 (achieved), and 此项 (this, this item) which are not found in the stoplist for traditional Chinese. No explanation has been documented for this phenomenon.

6.3 Querying

6.3.1 CONTAINS Operator

A query against documents that are indexed with a *context* index is issued by using the CONTAINS operator in the WHERE clause of a SELECT statement.

For example,

```
SELECT title from newsindex
WHERE CONTAINS (text, 'oracle') > 0; (6.5)
```

6.3.2 Scoring Algorithm

The `SELECT` statement in (6.5) may be modified by including a `SCORE()` operator to rank retrieved documents by their degree of relevance to the query.

```
SELECT title, SCORE(1) from newsindex
  WHERE CONTAINS (text, 'Oracle', 1) > 0
  ORDER BY SCORE(1) DESC;                                     (6.6)
```

The relevance score, `SCORE()`, uses an inverse frequency algorithm based on Salton's formula, where the relevance score for a word is proportional to

$$f (1 + \log(N/n)) \quad (6.7)$$

where F is the frequency of the search term in the document

N is the total number documents

n is the number of documents which contain the search term

Inverse frequency scoring assumes that the frequently occurring terms in a document set are noise terms, and so these terms are scored lower. For a document to score high, the query term must occur frequently in the document but infrequently in the document set as a whole (*Oracle Corp. (2003)*).

6.3.3 CONTAINS Query Operators

A number of query operators are supported within the `CONTAINS` operator. These query operators can be divided into four categories as shown in Table 6.4.

Binary Operators	Proximity and Location Operators	Expansion Operators	Theme Operators
AND (& OR () NOT (~) MINUS (-) EQUIValence (=) ACCUMulate (.) Weight (*)	NEAR (;) WITHIN	∇ Wildcards (%_) ∇ stem (\$) ∇ soundex (!) ∇ fuzzy (?)	∇ ABOUT Broader Term (BT, BTG, BTP, BTI) Narrower Term (NT, NTG, NTP, NTI) SYNonym (SYN) Translation Term (TR) Translation Term Synonym (TRSYN) Top Term (TT)

Table 6.4: Categorization of Query Operators

∇ not applicable in Chinese queries

Here are some examples on the use query operators:

(a) Binary Operators

Query Expression	Remarks
字符集 AND 中文 (character set) AND Chinese	<ul style="list-style-type: none"> retrieves documents that contain at least one occurrence of each of the query terms 字符集 and 中文 returns score of the lowest query term
二元结构 OR 切词单位 bi-gram OR token	<ul style="list-style-type: none"> retrieves documents that contain at least one occurrence of any of the query terms 二元结构 and 切词单位 returns score of the highest query term
文体分割 NOT 语义分析 (text segmentation) NOT (semantic analysis)	<ul style="list-style-type: none"> retrieves documents that contain 文体分割 but not 语义分析
文体分割 MINUS 语义分析 (text segmentation) MINUS (semantic analysis)	<ul style="list-style-type: none"> retrieves documents that contain 文体分割, and the presence of 语义分析 ranks a document lower calculates score by subtracting occurrences of 语义分析 from occurrences of 文体分割
二元结构 EQUIV 二元语义 bi-gram EQUIV (bigram semantics)	<ul style="list-style-type: none"> specifies 二元语义 as an acceptable substitution for 二元结构 returns score calculated as sum of occurrences of both query terms
文体分割, 切词单位, 语义分析 (text segmentation), token, (semantic analysis)	<ul style="list-style-type: none"> retrieves documents that contain at least one occurrence of any of the query terms 文体分割, 切词单位, or 语义分析 assigns the highest score to a document that contains all three terms 文体分割, 切词单位 and 语义分析
文体分割*2 AND 语义分析 (text segmentation)*2 AND (semantic analysis)	<ul style="list-style-type: none"> retrieves documents that contain both query terms 文体分割 and 语义分析 returns score that sums the occurrences of 语义分析 and twice the occurrences of 文体分割
文体分割*2, 切词单位, 语义分析 (text segmentation), token, (semantic analysis)	<ul style="list-style-type: none"> retrieves documents that contain at least one occurrence of any of the query terms 文体分割, 切词单位, or 语义分析 ranks documents according to document term weight, with the highest scores assigned to documents that have the highest total term weight e.g. a document containing 文体分割 and 切词单位 (total term weight = 3) is ranked higher than another that contains 切词单位 and 语义分析 (total term weight = 2)

(b) Proximity and Location Operators

Query Expression	Remarks
文体分割 NEAR 语义分析 (text segmentation) NEAR (semantic analysis)	<ul style="list-style-type: none"> retrieves documents that contain 文体分割 and 语义分析 occurring within the proximity of 100 tokens* returns higher scores for terms closer together and lower scores for terms farther apart in a document
NEAR((文体分割, 语义分析), 10) NEAR((text segmentation), (semantic analysis), 10)	<ul style="list-style-type: none"> retrieves documents that contain 文体分割 and 语义分析 occurring within the proximity of 10 tokens* returns higher scores for terms closer together and lower scores for terms further apart in a document
二元结构 WITHIN title bigram WITHIN title	<ul style="list-style-type: none"> searches 二元结构 within a pre-defined section "title"
(语义 AND 分析) WITHIN SENTENCE (semantics AND analysis) WITHIN SENTENCE	<ul style="list-style-type: none"> retrieves documents that contain both query terms 语义 and 分析 within one sentence
(语义 AND 分析) WITHIN PARAGRAPH (semantics AND analysis) WITHIN PARAGRAPH	<ul style="list-style-type: none"> retrieves documents that contain both query terms 语义 and 分析 within a paragraph

(c) Expansion Operators

The *wildcards* (%), *stem* (\$), *soundex* (!), and *fuzzy* operators are used to expand queries to include words of similar spelling, sound or of the same linguistic root. Such expansions are, however, not applicable to Chinese queries due to the fact that Chinese words are formed by character strings, and not letters.

* This is equivalent to proximity of 100 and 10 characters (respectively) if a *Chinese_Vgram_Lexer* is used. Reason: every character is an indexing point (see Section 6.2.2(a)).

(d) Theme Operators

The ABOUT theme operator interprets query terms using a supplied knowledge base. This feature is currently supported for English and French only.

The remaining theme operators such as broader term, narrower term, synonym, related term, etc. expand a query to include terms that has been defined in a thesaurus.

Consider a user-defined thesaurus *CHIN_THES* that contains the following relationships:

	(English equivalent of <i>CHIN_THES</i>)
低收入家庭	Low income families
UF 贫困家庭	UF Poor families
BT 社会问题	BT Social problems
贫困家庭	Poor families
PT 低收入家庭	PT Low income families
家庭问题	Family problems
BT 社会问题	BT Social problems
社会	Society
NT 社会问题	NT Social problems
社会福利团体	Charitable organizations
RT 社会问题	RT Social problems
NT 社会工作者	NT Social workers
社会工作者	Social workers
SYN 辅导员	SYN Counsellor
BT 社会福利团体	BT Charitable organizations
社会问题	Social problems
BT 社会	BT Society
NT 家庭问题	NT Family problems
NT 低收入家庭	NT Low income families
RT 社会福利团体	RT Charitable organizations

Then,

- NT (社会问题, 1, CHIN_THES) returns documents that contain 社会问题, 家庭问题, or 低收入家庭
- BT (社会问题, 1, CHIN_THES) returns documents that contain 社会问题 or 社会
- RT (社会问题, 1, CHIN_THES) returns documents that contain 社会问题 or 社会福利团体
- PT (贫困家庭, CHIN_THES) returns documents that contain 低收入家庭
- SYN (社会工作者, CHIN_THES) returns documents that contain 社会工作者 or 辅导员

7 Performance Evaluation of *Oracle® Text* in Chinese IR

7.1 Overview

7.1.1 Aims and Objectives

The aim of this chapter is to investigate the factors affecting Chinese IR, as well as to evaluate the performance of *Oracle® Text* for Chinese IR. Experiments were conducted using the framework of the TREC-5 Chinese Track, and 19 topics (CH1-CH16, CH21-CH23)¹⁵ were searched against a static set of approximately 170 megabyte of articles drawn from the *People's Daily* newspaper and the *Xinhua* newswire of Mainland China. Performance evaluation is based on the four measures used in TREC, namely:

- Summary Statistics
- Recall-Precision Averages
- Document Level Averages
- Average Precision Histogram

Individual experiments were first carried out to investigate the effects different parameters have on retrieval performance:

- Experiment I: Performance Comparison between *Chinese_Vgram_Lexer* and *Chinese_Lexer*
- Experiment II: Effect of Weighted Search Terms on Retrieval Performance
- Experiment III: Effect of Stopwords on Retrieval Performance
- Experiment IV: Effect of Thesaurus on Retrieval Performance
- Experiment V: Retrieval Performance for Short Queries

Finally, the best results obtained from the experiments were compared to those of TREC-5 participants to benchmark the retrieval performance of *Oracle® Text*.

¹⁵ Although 28 statements were provided in the TREC-5 Chinese Track, only 19 of them were actually evaluated in TREC-5 itself (see Appendix A).

7.1.2 Resources

The resources for the experiments were:

- Chinese newspaper corpus of TREC-5
- 19 topic statements (CH1-CH16, CH21-CH23)
- Relevance judgment of the 19 topics
- Results of the 10 participating groups in TREC-5 Chinese track
- Chinese thesaurus from the *Singapore Press Holdings* (SPH)

While the Chinese corpus was acquired from the *Linguistic Data Consortium* (LDC), the Chinese thesaurus was obtained from the *Singapore Press Holdings* (SPH) through *Digital Collections* in Germany and *Atex Media Command* in Singapore. The remaining resources were downloaded from the Data/Non-English section of the TREC Web site¹⁶.

Due to a difference in format, extensive editing was required for the approximately 1.3 megabyte thesaurus, to make it compatible to the *Oracle® Text* environment. Examples of some changes made are shown in Table 7.1.

Original Format	Revised Format
阿尔及利亚画家 / 艺术家 BT 阿尔及利亚人物	阿尔及利亚画家 SYN 阿尔及利亚艺术家 BT 阿尔及利亚人物
德国航空 — LUFTHANSA GERMAN AIRLINES BT 外国航空公司	德国航空 UF LUFTHANSA GERMAN AIRLINES BT 外国航空公司 LUFTHANSA GERMAN AIRLINES PT 德国航空
布什（美国总统） BUSH, GEORGE W BT 美国人物个人档 BU — BY	布什（美国总统） UF BUSH, GEORGE W BT 美国人物个人档 BU — BY BUSH, GEORGE W PT 布什（美国总统）
电话 / 随身电话 BT 电信配备 / 电讯器材	电话 SYN 随身电话 BT 电信配备 BT 电讯器材

Table 7.1: Revision of the SPH Thesaurus for Use in *Oracle® Text*

¹⁶ http://trec.nist.gov/data/docs_noneng.html

7.1.3 Experiment Environment

(a) System and Database setup:

All experiments were conducted using an *Oracle10g* (beta version) database that was run on a Linux operating system with a Chinese locale:

System Setup

Operating System	SuSE Linux Enterprise Server 8.0
XIM-Server for Chinese Input	XCIN version 2.5.2.3
Language/Locale Settings (in etc/sysconfig/language)	RC_LANG="zh_CN.GB18030" LANG="zh_CN.GB18030" LC_CTYPE="zh_CN.GB18030" LC_MESSAGES="zh_CN.GB18030"

Database Setup

Database	Oracle (10.1.0.1.0 Beta version)
NLS Character Set	UTF8
NLS National Character Set	UTF8
NLS Language	"SIMPLIFIED CHINESE_CHINA.ZHS32GB18030"

(b) Index Preferences

The articles of the TREC-5 Chinese corpus were imported into a table named *trec* which comprised the text columns *id* and *docs* (of `varchar2` and `CLOB` data types, respectively). The former stored document numbers and the latter the article contents.

Two separate *context* indices, namely *zh_index* and *ch_index*, were used in the experiments. These indices were created on the *docs* column using the preferences as shown in Table 7.2. A full description of the two indices can be found in Appendix B.

<i>Index</i>	<i>zh_index</i>	<i>ch_index</i>
<i>Nr. of unique tokens</i>	867,323	1,775,059
<i>Index Preferences</i>		
<i>Datastore</i>	CTXSYS.DEFAULT_DATASTORE (i.e. direct datastore)	
<i>Filter</i>	CTXSYS.NULL_FILTER	
<i>Lexer</i>	CHINESE_VGRAM_LEXER	CHINESE_LEXER
<i>Wordlist</i>	BASIC_WORDLIST <ul style="list-style-type: none"> ▪ STEMMER = NULL ▪ FUZZY_MATCH = CHINESE_VGRAM 	
<i>Storage</i>	CTXSYS.DEFAULT_STORAGE (i.e. basic storage)	
<i>Section Group</i>	BASIC_SECTION_GROUP <ul style="list-style-type: none"> ▪ Zone Sections: doc, text, headline, para, line ▪ Field Sections: docid, docno, date 	
<i>Stoptlist</i>	BASIC_STOPLIST <ul style="list-style-type: none"> ▪ Default stoptlist for simplified Chinese 	

Table 7.2: A Comparison of *zh_index* and *ch_index*

7.2 Experiments

This section summarizes the results and observations of all the experiments conducted. The search statements and detailed results (consisting of the four TREC evaluation measures) of the individual experiments are found in Appendix C and D, respectively.

7.2.1 Experiment I: Performance Comparison between *Chinese_Vgram_Lexer* and *Chinese_Lexer*

Procedure

The performance comparison between the *Chinese_Vgram_Lexer* and *Chinese_Lexer* was conducted using both Automatic and Manual runs. The search statements for the runs were constructed using the following guidelines:

- (i) Automatic run: combine all descriptors in a topic with the ACCUM operator
- (ii) Manual run: combine all descriptors in a topic with Boolean, ACCUM or Weight operators, and revise queries after examining the retrieved documents through adding new terms or eliminating existing terms

The ACCUM (accumulate) operator is ideal for TREC queries because it behaves like the OR operator, except that it assigns higher ranks to documents that match more terms and terms with higher weights (Mahesh/Kud/Dixon (1999)). The Boolean operators, on the other hand, are good for refining queries to combine synonyms (OR) or joining important concepts (AND).

Example:

Query terms:

新疆 (Xinjiang) }
 维吾尔 (Uygur) } synonyms
 边境贸易 (border trade) }
 边贸 (abbreviation of "border trade") } synonyms

Query statement:

→ (新疆 OR 维吾尔)
 AND (边境贸易 OR 边贸)

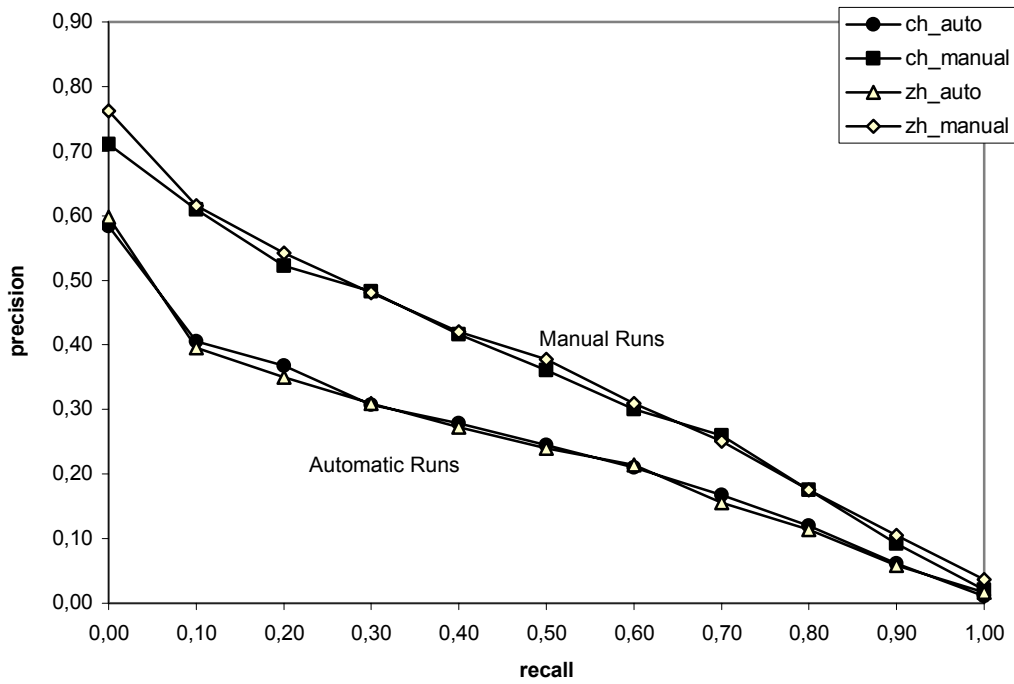
The search statements for all 19 topics (see Appendix C) were run against the TREC collection, and four sets of results were obtained (see Table 7.3).

Results:

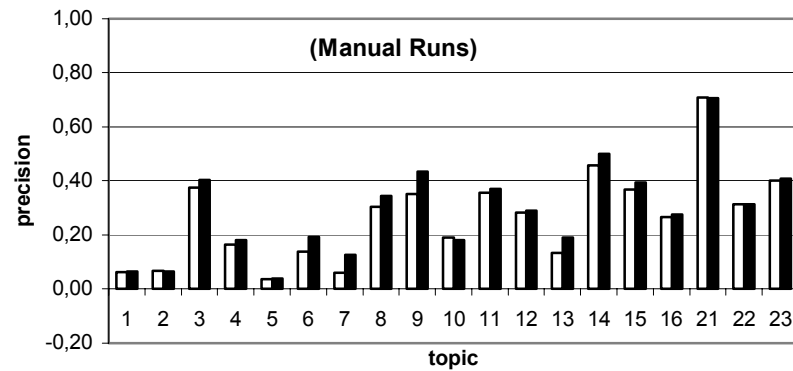
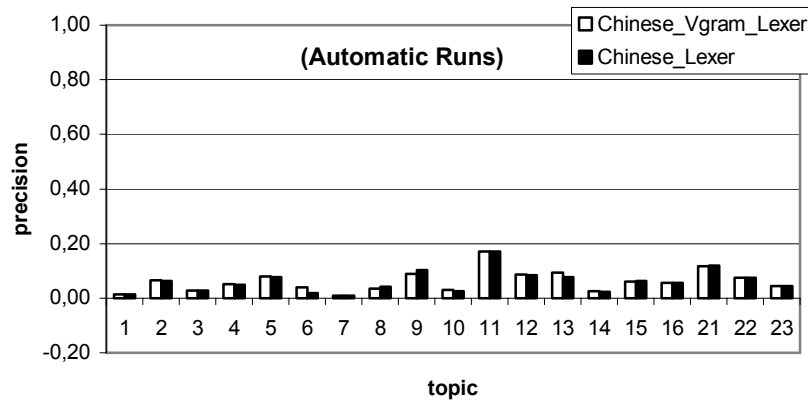
<i>Run</i>	Automatic		Manual	
<i>Result Set</i>	ch_auto	zh_auto	ch_manual	zh_manual
<i>Search Settings</i>				
<i>Nr. of topics searched</i>	19		19	
<i>Index</i>	<i>ch_index</i>	<i>zh_index</i>	<i>Ch_index</i>	<i>zh_index</i>
<i>Summary of Search Results</i>				
<i>Nr. of docs retrieved</i>	17502	17503	4280	5229
<i>Nr. of docs relevant</i>	1027	1055	951	1012
<i>Overall Recall</i>	0.7341	0.7541	0.6798	0.7234
<i>Average Precision</i>	0.2863	0.2728	0.4706	0.4594
<i>R-Precision</i>	0.2886	0.2888	0.4205	0.4176

Table 7.3: Summary of Search Results for Experiment I

Recall-Precision Curve:



Precision Histograms¹⁷:



¹⁷ Note that comparisons in these histograms are made on exact precision values, instead of the median precision values.

Observations:*(a) Automatic versus Manual runs*▪ Average Precision and R-Precision Values

Better precision was obtained in the Manual runs, namely result sets “zh_manual” and “ch_manual”, where queries were manually formulated to include synonyms and to exclude unimportant words or noise in the search topics. The overall improvements in the average precision and R-precision values (as compared to the Automatic runs) were over 60% and 40%, respectively.

▪ Recall

The overall recall, on the other hand, was better in the Automatic runs (result sets “zh_auto” and “ch_auto”), where only the ACCUM operator was used to connect descriptors provided in the topic statements.

(b) Chinese_Lexer versus Chinese_Vgram_Lexer▪ Average Precision and R-Precision Values

The choice of lexer had produced little effect on the average precision and R-precision values. While the average precision was slightly better in the result sets “ch_auto” and “ch_manual” where the *Chinese_Lexer* was used (4.9% and 2.9%, respectively), the R-precision remained almost constant.

▪ Automatic Runs

The precision histograms for the Automatic runs showed that the performance of individual search topics was not affected by the lexer choice. When the result sets “ch_auto” and “zh_auto” were compared, 10 topics had identical results, and the difference in precision readings for the remaining nine topics was 0.022 or less.

▪ Manual Runs

The *Chinese_Lexer* was found to produce better results for the Manual runs. As shown in Figure 7.1, 15 topics performed better in the result set “ch_manual”. On the average, the precision values of the individual topics were 14.9% better in “ch_manual”. The best improvement in precision was in topic 7 “Claims made by both PRC and Taiwan over islands in the South China Sea” (中国大陆与台湾对南海诸岛的立场), where an improvement of over 100% was observed. This topic was also found to be the longest query, whereby a total of 15 search terms was used. Generally, it was observed that when more search terms were

present in a query, the difference between the precision values of “ch_manual” and “zh_manual” for this query would also be more significant (see Figure 7.1).

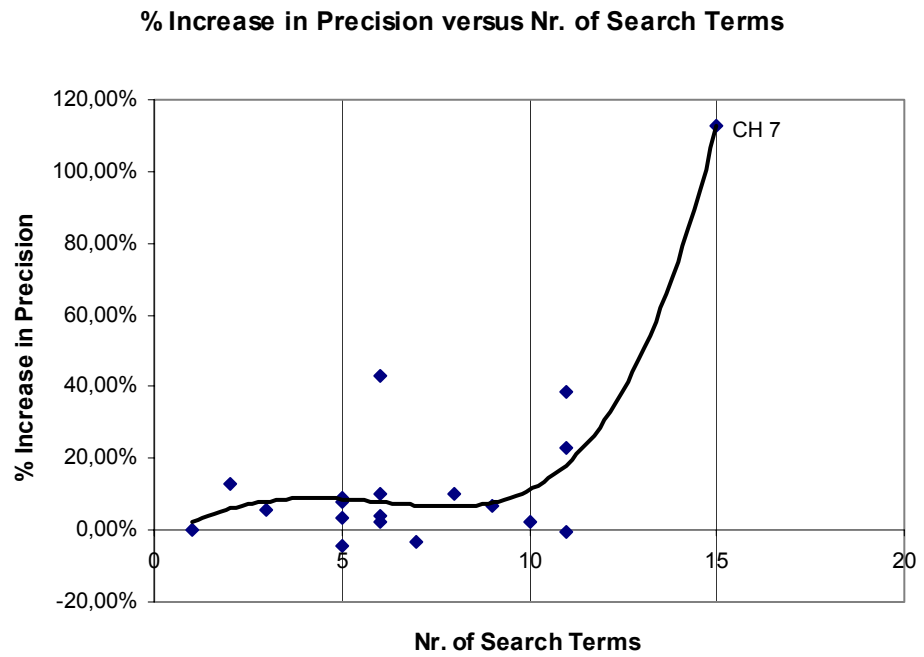


Figure 7.1: Percentage Increase in Precision Values Between “ch_manual” And “zh_manual” Across All Topics

(c) *Issues in query formulation*

- Transliteration

In topic CH14 “Cases of AIDs in China”, the TREC-5 topic description used the word 爱滋病 for *AIDs*. This word is more commonly used in Hong Kong and Taiwan, but not Mainland China. A search using the term 爱滋病 on the TREC-5 corpus, which is a collection of articles from China’s *Peoples’ Daily* newspapers and *Xinhua* news agency, retrieved less than 5 relevant documents. It is therefore necessary to search using the term 艾滋病, a transliteration for *AIDs* used in Mainland China.

- Abbreviated noun-phrases

In topic CH12, 世界妇女大会 (World Conference on Women) was abbreviated to 世妇会 in most of the relevant documents. Failing to include the abbreviation as a synonym in the query significantly reduced the recall rates.

- “Real words” versus Parts of Words

In topic CH15, the word 海地 (Haiti) appeared not only as a country name in the corpus, but also as parts of words or phrases, like in 上海地下党 (underground organizations in Shanghai), 沿海地区 (coastal regions), etc. In all 295 documents which were indexed by *ch_index* (i.e. with lexer preference set as *Chinese_Lexer*) to contain the token 海地, the country name “Haiti” was indexed correctly. In comparison, *zh_index* indexed 1,167 documents with the token 海地, whereby the majority of these documents used 海地 as parts of words or phrases only.

- Context-related Words

Using the term 中国 (Mainland China) in combination with an AND operator to limit a search within the Chinese context had sometimes caused recall rates to suffer. Reason being, the terms 我国 (our country), 我们 (we), 祖国 (homeland), etc. were often used affectionately in the documents to refer to “Mainland China”.

7.2.2 Experiment II: Effect of Weighted Search Terms on Retrieval Performance

Procedure:

This experiment investigated the effect of weighted search terms on retrieval performance in an Automatic Run. The *ch_index* (i.e. *Chinese_Lexer* as lexer preference) was used, and a new set of queries with weighted terms was created and run against the corpus. The new result set, referred to as “weighted_auto”, was compared with that of a control set, namely, “ch_auto” of Experiment I.

The new queries were formulated using the following simple algorithm:

- (i) Combine all descriptors in a topic by the ACCUM operator
- (ii) Increase the weight of each descriptor based on its frequency of occurrence in the entire topic description, i.e. including title and narrative

Example:

Consider the topic statement shown in Figure 5.6 of Chapter 5, where the frequencies of occurrence of the individual descriptors are as follows:

<i>Descriptor</i>	<i>Freq of Occurrence</i>
最惠国待遇	4
中国	2
人权	1
经济制裁	1
分离	1
脱钩	1

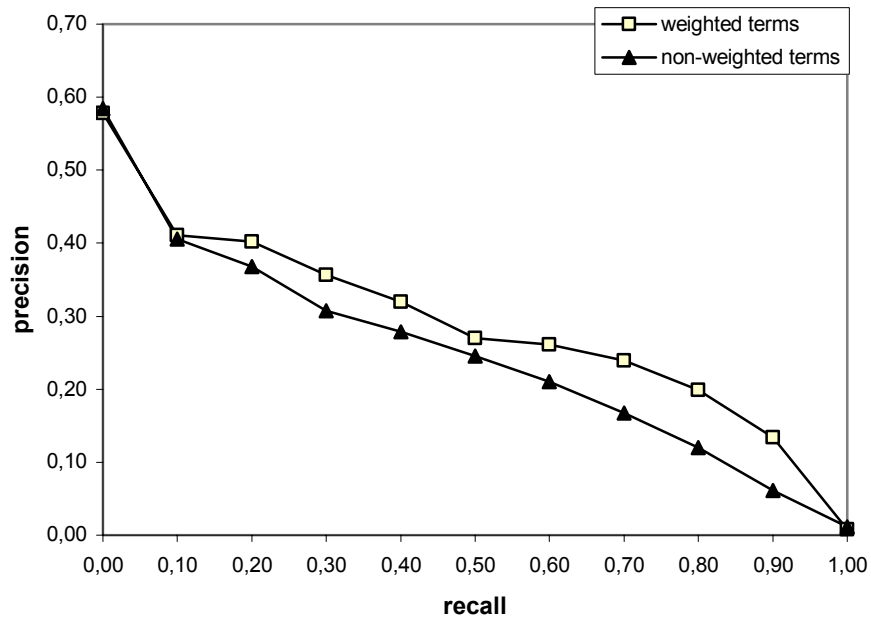
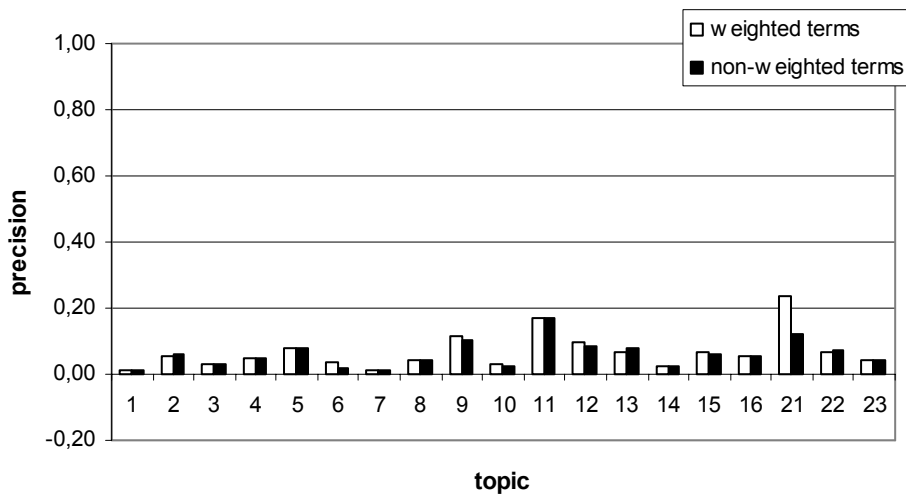
The new query corresponding to this topic is therefore:

最惠国待遇*4, 中国*2, 人权, 经济制裁, 分离, 脱钩

Results:

<i>Result Set</i>	weighted_auto	ch_auto
<i>Search Settings</i>		
<i>Run</i>	Automatic	
<i>Nr. of topics searched</i>	19	
<i>Index</i>	<i>ch_index</i>	
<i>Summary of Search Results</i>		
<i>Nr. of docs retrieved</i>	17500	17502
<i>Nr. of docs relevant</i>	1175	1027
<i>Overall Recall</i>	0.8399	0.7341
<i>Average Precision</i>	0.3309	0.2863
<i>R-Precision</i>	0.3327	0.2886

Table 7.4: Summary of Search Results for Experiment II

Recall-Precision Curve:Precision Histogram:Observations:

- As summarized in Table 7.5 on next page, the use of weighted terms in queries improved the overall retrieval performance. On average, each topic had over 10% improvement in its precision and recall values, as well as 44% increase in R-precision value, when weighted terms were used.

- The best improvement was achieved for topics CH6 and CH21, whereby both precision and R-precision values increased over 87%. This observation could be related to the fact that the important keywords in these search statements, namely 世界贸易组织 (World Trade Organization (WTO)) and 彭定康 (Peng Ding-Kang, the former governor of Hong Kong), were given significantly higher relative weights:

CH6: 世界贸易组织*3, 关贸总协, 市场准入, 世界贸易体系, 多边贸易, 成员
 CH21: 香港问题*2, 特别行政区, 彭定康*7, 计划, 建议

- A drop in precision values was observed in topics CH2 and CH13, when the relative weight of the term 中国 (Mainland China) was raised.

CH2: 中国*2, 一国两制*2, 台湾*3, 和平统一, 经贸合作, 两岸关系, 科技*2, 文化*2, 交流*2
 CH13: 中国*4, 经济实力, 奥运*4, 世界运动大会, 奥林匹克, 筹备工作

Topic	Precision Values			R-Precision Values		
	weighted_auto	ch_auto	% Diff	weighted_auto	ch_auto	% Diff
CH1	0.0130	0.0130	0.00%	0.0000	0.0000	0.00%
CH2	0.0560	0.0630	-11.11%	0.3188	0.3188	0.00%
CH3	0.0280	0.0280	0.00%	0.5517	0.5517	0.00%
CH4	0.0500	0.0500	0.00%	0.3333	0.3137	6.25%
CH5	0.0769	0.0767	0.33%	0.1071	0.1071	0.00%
CH6	0.0390	0.0180	116.67%	0.1948	0.1039	87.50%
CH7	0.0100	0.0100	0.00%	0.2353	0.1176	100.00%
CH8	0.0420	0.0410	2.44%	0.3256	0.3256	0.00%
CH9	0.1130	0.1030	9.71%	0.3852	0.2951	30.56%
CH10	0.0330	0.0250	32.00%	0.2857	0.1429	100.00%
CH11	0.1670	0.1700	-1.76%	0.2742	0.4194	-34.62%
CH12	0.0940	0.0840	11.90%	0.3193	0.3361	-5.00%
CH13	0.0660	0.0780	-15.38%	0.2273	0.2727	-16.67%
CH14	0.0230	0.0230	0.00%	0.1754	0.0351	400.00%
CH15	0.0680	0.0630	7.94%	0.4638	0.2899	60.00%
CH16	0.0560	0.0560	0.00%	0.2241	0.2414	-7.14%
CH21	0.2350	0.1200	95.83%	0.8319	0.3235	157.14%
CH22	0.0697	0.0743	-6.20%	0.6667	0.7333	-9.09%
CH23	0.0450	0.0440	2.15%	0.4000	0.5556	-28.00%
% Ave increase in Precision			12.88%	% Ave increase in R-Precision		44.26%

Table 7.5: Comparison of Average Precision and R-precision Values in Experiment II

7.2.3 Experiment III: Effect of Stopwords on Retrieval Performance

Procedure:

This experiment compared the performance for 3 different stoplist settings:

- Without stoplist
- Oracle 's default stoplist for simplified Chinese¹⁸
- Modified stoplist containing 23 most frequently occurring functional words in the corpus (with document frequency between 19,000 and 124,000)¹⁹

A total of six sets of readings were taken based on the different variable settings, as shown in Table 7.6.

<i>Result Set</i>	<i>Stoplist</i>	<i>Run</i>	<i>Index</i>	<i>Query Set</i>
Stop1_auto	Without stoplist	Automatic	<i>ch_index</i>	Query set of Automatic Run in Experiment I
Stop2_auto	Oracle's default stoplist for simplified Chinese			
Stop3_auto	Modified stoplist containing 23 most frequently occurring functional words in the corpus			
Stop1_manual	As in "Stop1_auto"	Manual		Query set of Manual Run in Experiment I
Stop2_manual	As in "Stop2_auto"			
Stop3_manual	As in "Stop3_auto"			

Table 7.6: Run and Stoplist specification for the six sets of readings

Results and Observations:

- (a) The number of unique tokens remained almost unchanged for the different stoplists:

<i>Stoplist</i>	<i>Nr. of unique tokens</i>
Without stoplist	867,364
Oracle's default stoplist for simplified Chinese	867,323
Modified stoplist	867,375

- (b) Stopwords were found to have no effect on the search results: all Automatic runs (namely stop1_auto, stop2_auto and stop3_auto) produced identical results; likewise, the results of all Manual runs (stop1_manual, stop2_manual and stop3_manual) were the same.

¹⁸ See Appendix E

¹⁹ *ibid.*

7.2.4 Experiment IV: Effect of Thesaurus on Retrieval Performance

Procedure:

This experiment investigated whether the use of a thesaurus to expand queries could improve search performance.

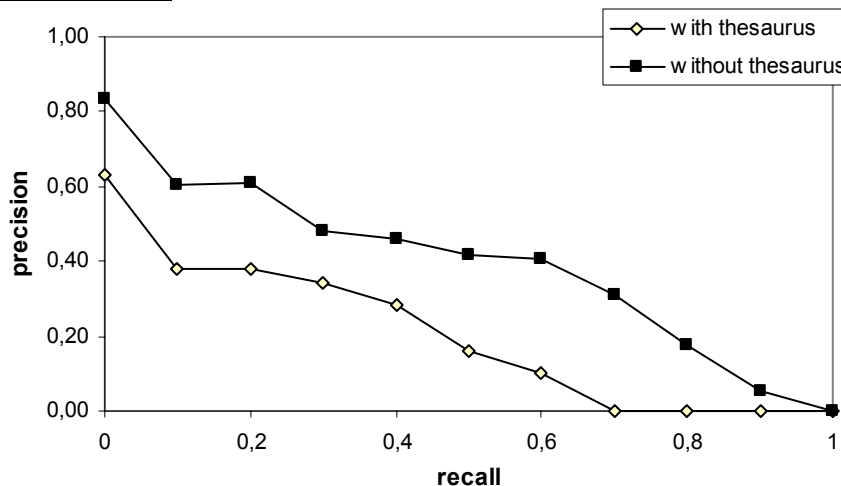
The SPH Chinese thesaurus was first edited to have a format suitable for *Oracle® Text*, and an attempt was made to select suitable thesaurus terms for the 19 TREC topics. However, no suitable terms were found for most topics, and the experiment was eventually conducted using topics CH3, 5, 13 and 23 only. The new result set “zh_thes” was compared with the result set “zh_manual” of Experiment I.

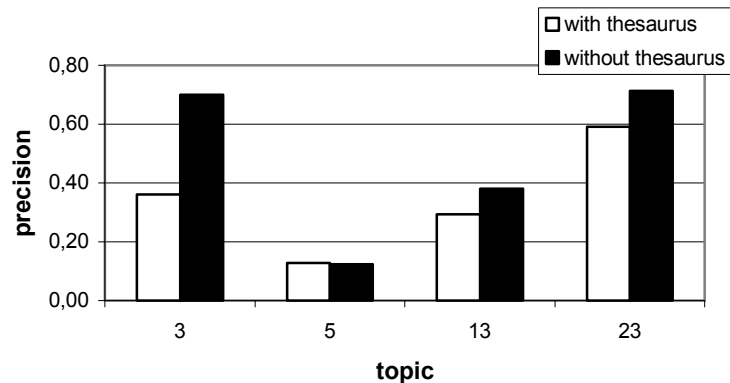
Results:

Result Set	zh_thes	zh_manual
<i>Search Settings</i>		
Run	Manual	
Nr. of topics searched	4	
Index	zh_index	
<i>Summary of Search Results</i>		
Nr. of docs retrieved	1178	1394
Nr. of docs relevant	122	179
Overall Recall	0.5728	0.8404
Average Precision	0.3436	0.4792
R-Precision	0.2934	0.4341

Table 7.7: Summary of Search Results for Experiment IV

Recall-Precision Curve:



Precision Histogram:Results and Observations:

- (a) No improvement in performance was achieved by including thesaurus terms in the queries for topics CH3, 5, 13 and 23. In fact, both precision and recall values were lower when the SPH thesaurus was used.
- (b) The SPH thesaurus was found to contain little meaningful synonyms, related terms or narrower terms to expand the TREC queries. Most terms in the thesaurus were person names or very broad classification of themes. For example:

中国	China
UF CHINA	UF CHINA
BT 国家 C—F	BT Countries C—F
NT 中国城市	NT Cities of China
NT 中国岛屿	NT Islands of China
NT 中国国庆	NT National Day of China
NT 中国环境	NT Environment of China
NT 中国环保	NT Recycling in China
NT 中国建设 (基础设施)	NT Buildings in China (Basic infrastructure)
NT 中国交通	NT Road Traffics in China
NT 中国教育	NT China' s Education
NT 中国节日	NT Festivals in China
NT 中国经济	NT China' s Economy
.	.
.	.
.	.

- (c) The terms defined in the SPH thesaurus were also biased towards the Singaporean or Southeast Asian context.
- (d) Due to the small number of queries used, and the mismatch between thesaurus terms and the vocabulary of the corpus, the hypothesis that search performance could be improved through the use of a thesaurus could not be supported.

7.2.5 Experiment V: Retrieval Performance for Short Queries

Procedure:

Long queries were used in the Manual runs of the previous experiments. On the average, each of these queries contained as many as seven search terms. Such long queries are considered unrealistic because “real-life queries are usually very short, like one or two words” (Kwok (1999)). The aim of this experiment is therefore to study the retrieval performance for short queries created by real users.

Six participants were selected in this experiment:

- User 1: music undergraduate student
- User 2 : graduate student of information and communication studies
- User 3 : high school physics teacher
- User 4 : primary school Chinese teacher
- User 5 : former news reporter for a Chinese radio station in Singapore
- User 6 : university research-assistant in the faculty of mechanical engineering

All participants are Chinese native speakers, and have little to extensive experience in searching the World Wide Web for information in the Chinese language. User 2, 3 and 5 are degree holders of Chinese studies, and User 5 is also experienced in searching the Chinese newspapers database of SPH.

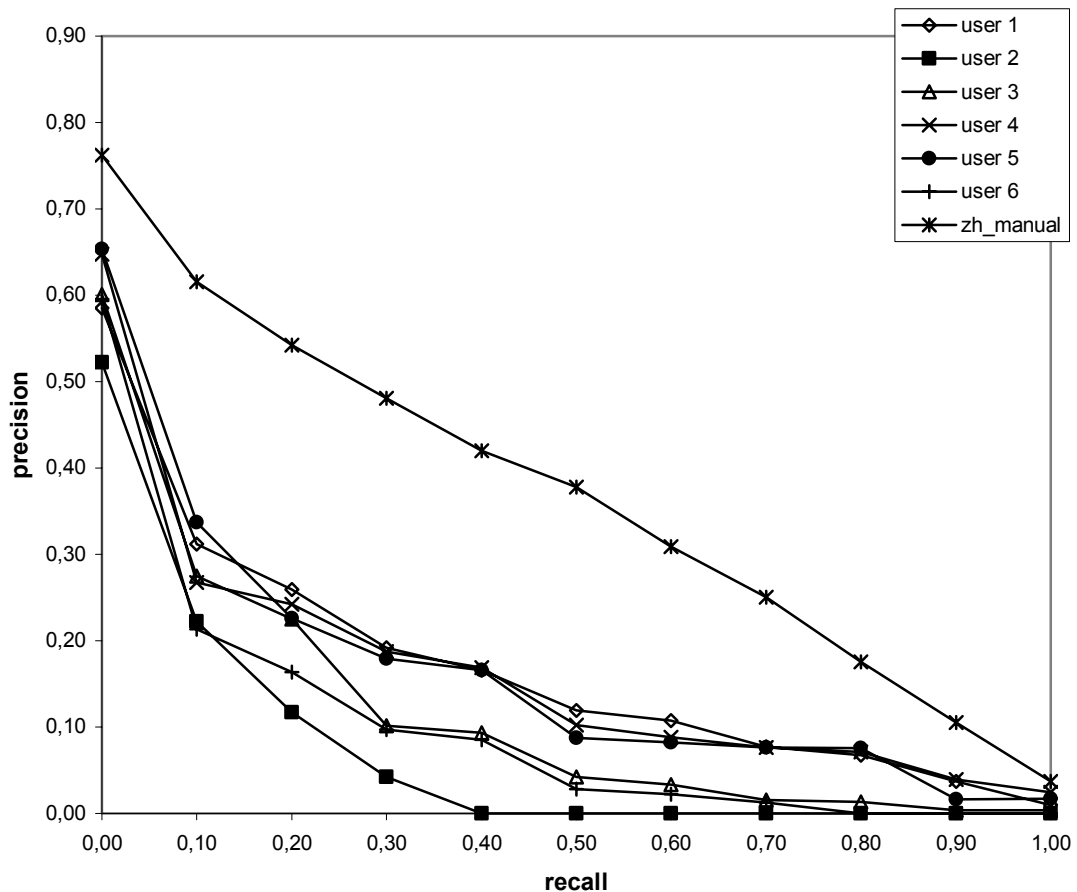
In the experiment, each participant was presented only the titles and narratives of the 19 TREC topics (i.e. without descriptor lists), and was asked to form query statements using either his/her own words, or the terms found in the titles and narratives. The purpose of hiding the descriptors from the participants was to avoid influence on the participants’ choices of query terms, and to prevent the formulation of lengthy queries.

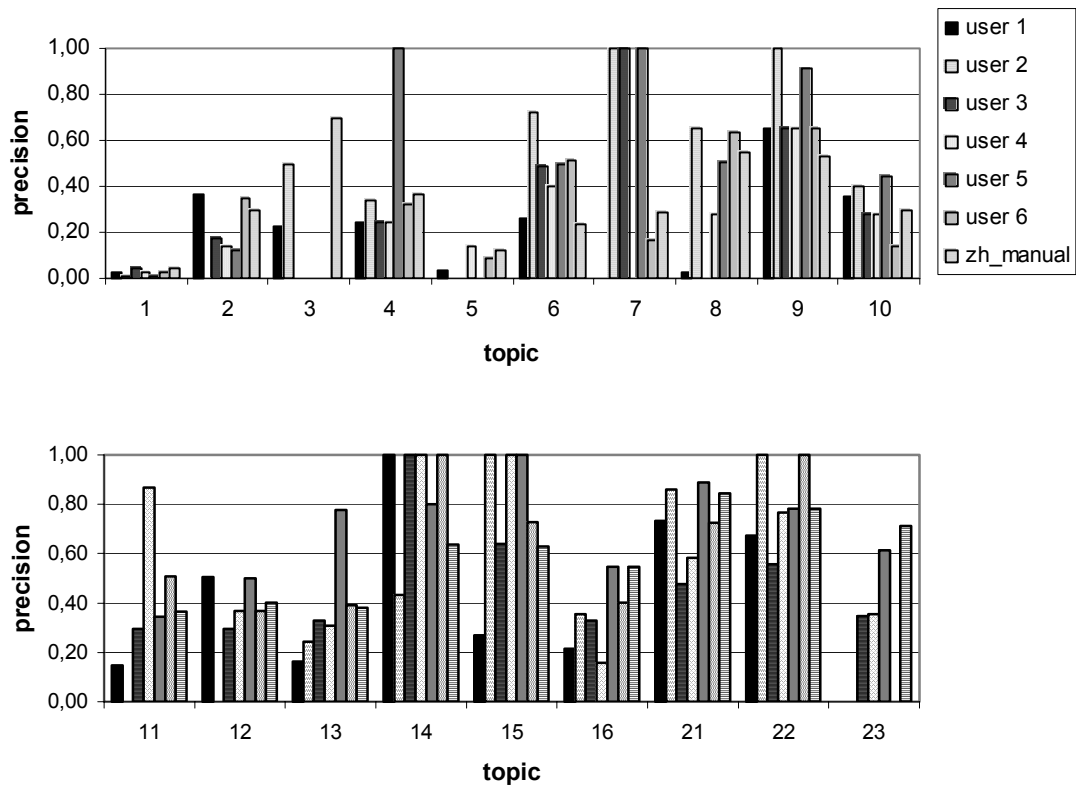
The user-formulated queries were then collected and searched against the Chinese corpus. As the participants were not directly involved in the actual search process, they were neither given any immediate feedback on the performance of their queries, nor allowed to revise their queries based on the search results.

Results:

Result Set	User 1	User 2	User 3	User 4	User 5	User 6	zh_manual
<i>Search Settings</i>							
Run	Manual						
Nr. of topics searched	19						
Index	zh_index						
<i>Summary of Search Results</i>							
Nr. of docs retrieved	11120	503	3739	8314	1570	1337	5229
Nr. of docs relevant	739	152	435	458	260	303	1012
Overall Recall	0.5282	0.1086	0.3109	0.2988	0.1858	0.2166	0.7234
Average Precision	0.3105	0.4480	0.3762	0.3979	0.5649	0.4214	0.4594
R-Precision	0.2013	0.0950	0.1690	0.1951	0.1822	0.1396	0.4176

Table 7.8: Summary of Search Results for Experiment V

Recall-Precision Curve:

Precision Histogram:Observations:*(a) Short Queries versus Long Queries*

- Query Length and Word Length

In comparison to the long queries in the result set “zh_manual”, the queries formed by the participants were relatively short; consisting of only two to three query terms on the average (see Appendix C). In addition, half of the participants, namely User 1, User 2 and User 5, have searched more topics using long words that were between six to nine characters long (see Table 7.9). Examples of such words are 中国毒品问题, 香港立法改革 (six-character long), 妇女的社会地位 (seven-character long), and 第四届世界妇女大会 (nine-character long).

<i>Participant</i>	User 1	User 2	User 3	User 4	User 5	User 6	zh_manual
6-character	2	6	2	2	1	2	2
7-character		2					
9-character	1				2		
Total nr. of queries with words ≥ 6 characters	3	8	2	2	3	2	2

Table 7.9: Number of Queries Containing Words with Six or More Characters

- Precision and Recall

The recall rates were significantly lower for the user-formed queries (average 0.2748), and the R-precision readings ranged only between 0.0950 and 0.2013, as compared to a high 0.4176 of the “zh_manual” result set.

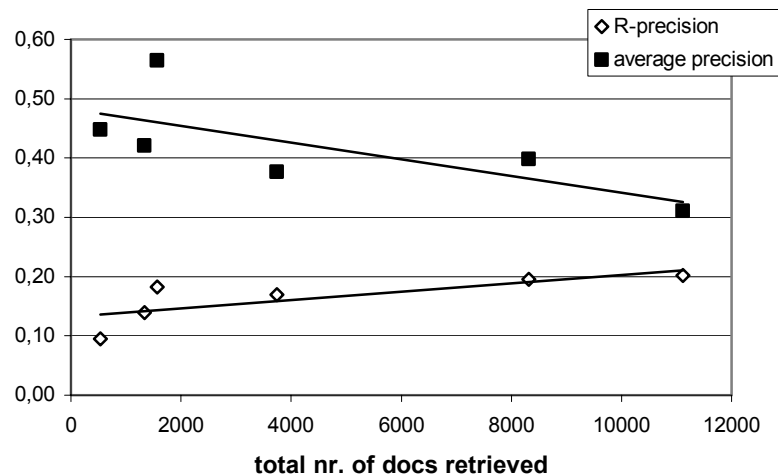
On the other hand, the average precision readings of the six user-formed queries (ranging from 0.3105 to 0.5649; average 0.4198) were comparable to that of “zh_manual” (0.4594).

In general, queries formed by long words consisting of six or more characters produced very high precision but extremely low recall.

(b) *Comparison of User-formed Queries*

- Variations of Average Precision and R-Precision

The total number of documents retrieved by each participant varied greatly between 503 and 11,120. While a steady drop in average precision was observed when the total number of retrieved documents increased, a more gradual increase in the R-precision was observed.



- Search Techniques

The most popular query operator was AND, followed by OR, ACCUM, Weight (*) and NOT. The NEAR operator was only used by User 5 to formulate queries. Although in general, different query operators performed well for different queries, and each participant had the best precision values for at least two topics (see Table 7.10 on next page), the NEAR operator was observed to produce the best overall precision at a reasonable recall. The Weight (*) operator was also found to improve relevance ranking and precision.

<i>Participant</i>	User 1	User 2	User 3	User 4	User 5	User 6
CH1	X			X		X
CH2	X					
CH3		X				
CH4					X	
CH5				X		
CH6		X				
CH7		X	X		X	
CH8						
CH9		X				
CH10						
CH11				X		
CH12	X					
CH13						
CH14	X		X	X		X
CH15						
CH16						
CH21					X	
CH22		X				
CH23					X	X
Nr. of best average precision	4	5	2	4	4	3

Table 7.10: Distribution of Best Precision Values

7.3 Benchmarking Performance against TREC-5 Results

The aim of this section is to benchmark the performance of *Oracle® Text* for Chinese IR. Benchmarking was carried out through a comparative study between the best results obtained from the experiments in Section 7.2, and the results of the nine participants of the TREC-5 Chinese Track.

7.3.1 TREC-5 Result Sets

A total of 20 result sets (or runs) were submitted by nine participating groups for the TREC-5 Chinese Track. These results are compiled in the conference proceedings of TREC-5, and are also available online at the TREC Homepage²⁰.

For the purpose of benchmarking *Oracle® Text*, only the best result set from each participant in each run was taken (i.e. nine in the Automatic Run; and four in the Manual Run). The best result sets, namely “weighted_auto” and “ch_manual”, were chosen from Section 7.2 to represent *Oracle® Text* in the Automatic Run and Manual Run, respectively.

²⁰ http://trec.nist.gov/pubs/trec5/t5_proceedings.html

<i>Participant</i>	<i>Nr. of Runs submitted</i>	
	Automatic	Manual
1) City University	2	0
2) Claritech Corporation	1	1
3) Cornell University	2	0
4) George Mason University	2	2
5) Information Technology Institute	1	1
6) Royal Melbourne Institute of Technology, RMIT	2	0
7) Queens College, CUNY	2	0
8) University of California, Berkeley	1	1
9) University of Massachusetts	2	0
Total Nr. of Runs	15	5

Table 7.11: Total Number of Result Sets for the Automatic and Manual Runs in the TREC-5 Chinese Track

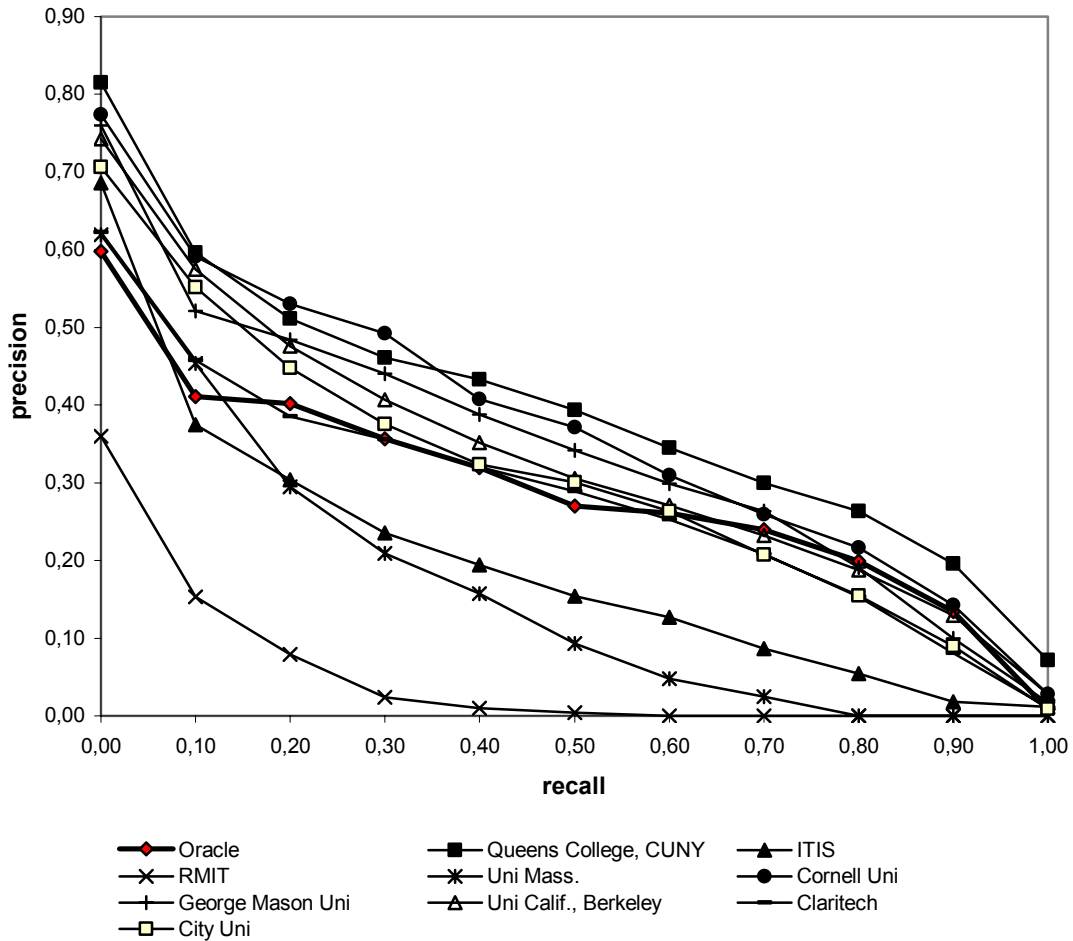
7.3.2 Comparative Study – Automatic Run

Summary Statistics:

<i>Result Set</i>	Oracle	Queens College, CUNY	ITIS	RMIT	Uni Mass.	Cornell Uni	George Mason Uni	Uni Calif., Berkeley	Claritech	City Uni
<i>Run</i>	Automatic									
<i>Nr. of topics searched</i>	19									
<i>Nr. of docs retrieved</i>	17500	19000	17501	18999	19000	19000	19000	19000	19000	19000
<i>Nr. of docs relevant</i>	1175	1313	899	360	542	1343	1250	1246	1182	1203
<i>Overall Recall</i>	0.8399	0.9385	0.6426	0.2573	0.3874	0.9600	0.8935	0.8906	0.8449	0.8599
<i>Average Precision</i>	0.3309	0.3789	0.1731	0.0371	0.1519	0.3598	0.3274	0.3192	0.2677	0.2943
<i>R-Precision</i>	0.3327	0.3823	0.2289	0.1045	0.2333	0.3829	0.3571	0.3565	0.2998	0.3050

Table 7.12: Summary of TREC-5 Search Results (Automatic)

Recall-Precision Curve:



Recall	Oracle	Queens College, CUNY	ITIS	RMIT	Uni Mass.	Cornell Uni	George Mason Uni	Uni Calif., Berkeley	Claritech	City Uni
0.00	0.5975	0.8150	0.6861	0.3597	0.6195	0.7744	0.7600	0.7429	0.6223	0.7063
0.10	0.4108	0.5958	0.3744	0.1534	0.4539	0.5910	0.5210	0.5745	0.4576	0.5513
0.20	0.4021	0.5116	0.3039	0.0790	0.2944	0.5302	0.4838	0.4758	0.3854	0.4477
0.30	0.3568	0.4614	0.2355	0.0241	0.2094	0.4922	0.4407	0.4068	0.3561	0.3753
0.40	0.3195	0.4335	0.1945	0.0100	0.1570	0.4079	0.3875	0.3513	0.3206	0.3239
0.50	0.2702	0.3940	0.1543	0.0043	0.0929	0.3712	0.3419	0.3051	0.2888	0.3007
0.60	0.2611	0.3453	0.1270	0.0000	0.0475	0.3100	0.2992	0.2709	0.2527	0.2639
0.70	0.2395	0.2998	0.0863	0.0000	0.0250	0.2594	0.2632	0.2326	0.2086	0.2078
0.80	0.1992	0.2637	0.0541	0.0000	0.0000	0.2165	0.1911	0.1874	0.1533	0.1546
0.90	0.1344	0.1957	0.0185	0.0000	0.0000	0.1421	0.1000	0.1293	0.0807	0.0897
1.00	0.0083	0.0714	0.0119	0.0000	0.0000	0.0284	0.0171	0.0280	0.0109	0.0087

Table 7.13: TREC-5 Recall-Precision Values (Automatic)

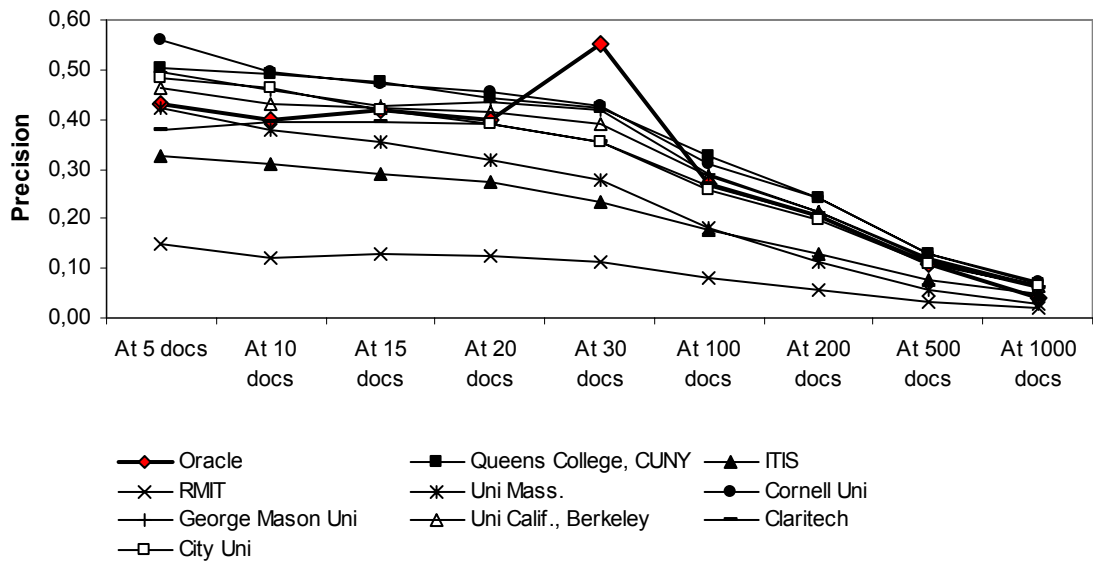
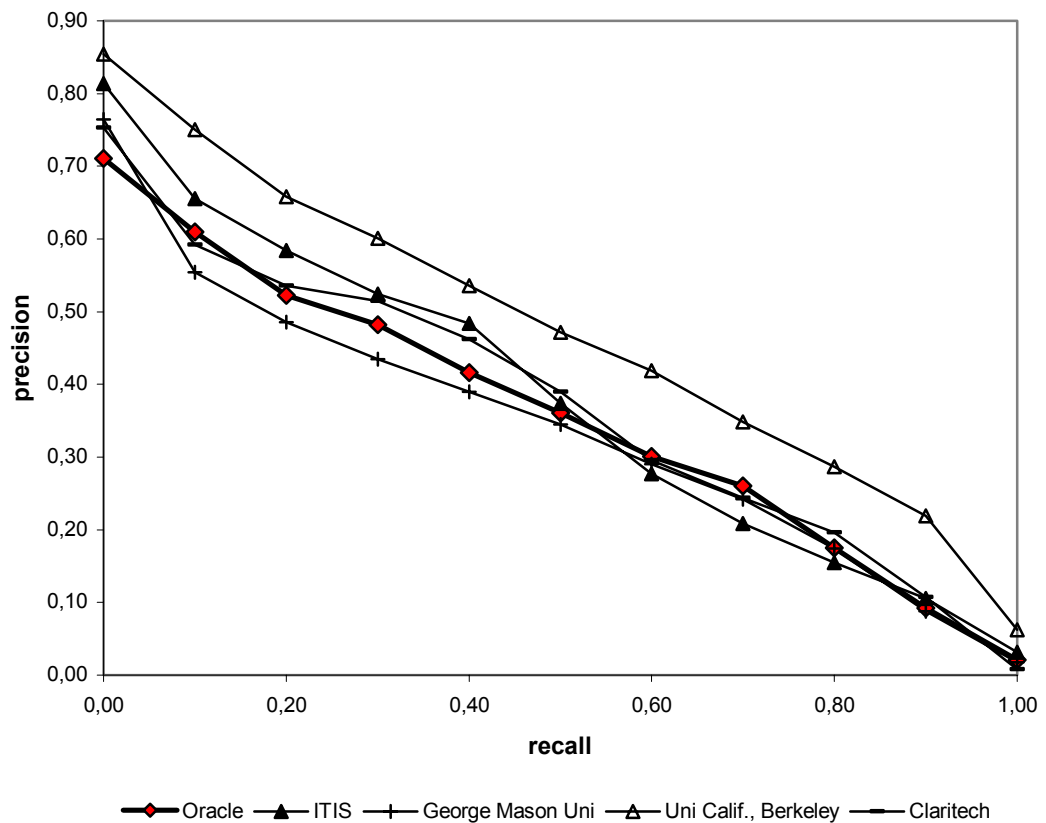
Document Level Precision:

Table 7.14: TREC-5 Document Level Precision Values (Automatic)

7.3.3 Comparative Study – Manual RunSummary Statistics:

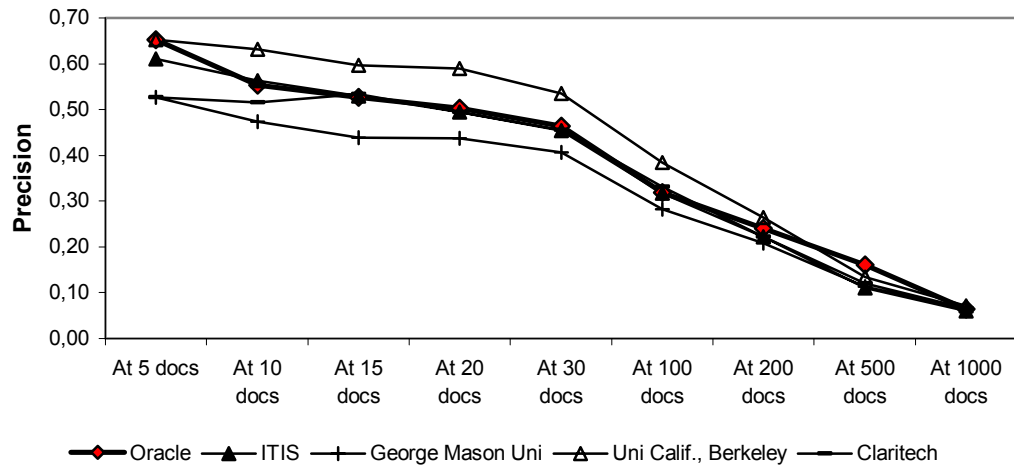
Result Set	Oracle	IT IS	George Mason Uni	Uni Calif., Berkeley	Claritech
Run	Manual				
Nr. of topics searched	19				
Nr. of docs retrieved	4280	19000	19000	19000	19000
Nr. of docs relevant	951	1136	1234	1364	1253
Overall Recall	0.6798	0.8120	0.9216	0.9750	0.8956
Average Precision	0.4706	0.3607	0.3279	0.4610	0.3583
R-Precision	0.4205	0.3881	0.3594	0.4642	0.3872

Table 7.15: Summary of TREC-5 Search Results (Manual Run)

Recall-Precision Curve:

Recall	Oracle	IT IS	George Mason Uni	Uni Calif., Berkeley	Claritech
0.00	0.7108	0.8139	0.7647	0.8539	0.7534
0.10	0.6094	0.6557	0.5539	0.7507	0.5923
0.20	0.5226	0.5845	0.4854	0.6584	0.5359
0.30	0.4825	0.5241	0.4349	0.6008	0.5150
0.40	0.4160	0.4837	0.3897	0.5354	0.4623
0.50	0.3609	0.3735	0.3451	0.4715	0.3893
0.60	0.3006	0.2775	0.2901	0.4188	0.2952
0.70	0.2602	0.2082	0.2417	0.3486	0.2435
0.80	0.1753	0.1545	0.1743	0.2872	0.1960
0.90	0.0923	0.1058	0.0879	0.2194	0.1070
1.00	0.0208	0.0321	0.0173	0.0623	0.0077

Table 7.16: TREC-5 Recall-Precision Values (Manual)

Document Level Precision:

Result Set	Oracle	IT IS	George Mason Uni	Uni Calif., Berkeley	Claritech
At 5 docs	0.6526	0.6105	0.5263	0.6526	0.5263
At 10 docs	0.5526	0.5632	0.4737	0.6316	0.5158
At 15 docs	0.5263	0.5298	0.4386	0.5965	0.5333
At 20 docs	0.5026	0.4947	0.4368	0.5895	0.4947
At 30 docs	0.4632	0.4544	0.4070	0.5351	0.4561
At 100 docs	0.3179	0.3174	0.2826	0.3842	0.3316
At 200 docs	0.2400	0.2221	0.2084	0.2634	0.2226
At 500 docs	0.1610	0.1105	0.1138	0.1346	0.1204
At 1000 docs	0.0640	0.0598	0.0649	0.0718	0.0659

Table 7.17: TREC-5 Document Level Precision Values (Manual)

7.3.4 Results Analysis

Most participants of the Chinese track explored the use of words versus n-grams, and the methods of manually modifying queries (see Table 5.2 of Chapter 5). Generally, “the best n-gram approaches, including single character approaches were comparable to the best word based approaches”, and the use of term expansion was also found to improve retrieval performance, and most of the best results were achieved through manual modification of queries (Smeaton/ Wilkinson (1997)).

The best three results in the Automatic Run were obtained by *Queens College, CUNY*, followed by *Cornell University*, and *George Mason University*. It was observed that the indexing approaches used by these three participants were all different, namely combination of dictionary-based and statistical techniques for word segmentation (*Queens College, CUNY*), bi-grams indexing (*Cornell University*), and single character indexing (*George Mason University*).

In the Manual Run, *University of California, Berkeley* had produced significantly better results than all other participants, through extensive query modification by adding new words, changing term weights, and adding negative terms.

7.3.5 Benchmarking Results

Oracle® Text performed well in the Manual Run. It had the best overall average precision, as well as the second best R-precision and Document Level Precision readings, as compared the results of the TREC-5 participants. The performance in the Automatic Run was worse, and the rankings for the different evaluation measures were between third and seventh, as shown in Table 7.18.

<i>Run</i>	Automatic Run	Manual Run
<i>Average Precision</i>	3 rd	1 st
<i>R-Precision</i>	5 th	2 nd
<i>Recall-Precision Curve</i>	7 th	4 th
<i>Document Level Precision</i>	6 th	2 nd

Table 7.18: Ranking of *Oracle® Text* performance against TREC-5 Results

The overall recall of *Oracle® Text* in the Manual Run was relatively low (68%), as a result of the significantly lower number of documents being retrieved, as compared to those of the TREC-5 participants. In addition, the TREC's pooling method for relevance judgment could have missed relevant documents that were not retrieved or ranked high by the TREC participants, but were present in *Oracle® Text's* results.

A comparison on the performance for the individual topics was, however, not possible, since no information was available about the TREC-5 participants' precision values for the individual topics.

On the whole, *Oracle® Text's* retrieval performance for Chinese IR has reached the standards set by the TREC-5 participants.

7.4 Conclusion

Extensive experiments were performed on the large Chinese newspaper corpus with the 19 topics of TREC-5 Chinese Track. The effects which some parameters have on the retrieval effectiveness of Chinese documents were investigated, and the following conclusions could be drawn from the experiments:

- a) Retrieval performance depends not only on the effectiveness of a retrieval system, but more substantially on the ways which queries are formulated. It was observed that:
 - Queries that were manually formulated and revised after examining retrieved documents provided much better results than those that were automatically generated. Similarly, the use of extensive query expansion, as observed in the results University of California, Berkeley in the Manual Run of TREC-5, improved retrieval performance significantly.
 - Different standards of transliteration and the extensive use of abbreviated noun-phrases in the Chinese language were observed to affect the retrieval effectiveness of Chinese IR. Recall suffered if these factors were not taken into consideration in the process of query formulation.
 - Short queries (averaging three words) that were formed by long words (i.e. six characters or more) performed worse than long queries (six or more words) with short words. Queries formulated by real users were generally short, and produced results of relatively high precision but extremely low recall.

- b) The *Chinese_Lexer* (mixing lexical terms and bi-grams), in general, provides better retrieval results than the *Chinese_Vgram_Lexer* (overlapping bi-grams). As observed:
 - The *Chinese_Lexer* provided better average precision values, and retrieved documents that matched “real words” (e.g. 海地 as Haiti and not 海地 as part of the phrase 沿海地区 or “coastal regions”).
 - The *Chinese_Lexer* produced a smaller index, and therefore entailed less space costs, and could also shorten retrieval time.

- c) The ACCUM, NEAR and Weight (*) operators are especially good for Chinese IR.
- The ACCUM operator ensured a high recall and high ranking for documents that matched more terms with higher weights.
 - The proximity operator NEAR gave very good precision and a reasonable recall.
 - The Weight (*) operator was also found to improve relevance ranking and precision.
- d) The presence of stopwords does not affect performance.
- No change in retrieval results was observed when different stoplists were used.
 - A further literature review confirmed that the presence of stopwords does not contribute much noise to Chinese IR. On the other hand, accidental inclusion of a “crucial” word in a stoplist may also harm retrieval substantially (Kwok (1999); Kwok/Grunfeld/Xu (1997); Nie/Brisebois/Ren (1996)).
- e) Query expansion using a thesaurus has not been proven to improve performance.
- This could be due to the mismatch between the thesaurus terms and the vocabulary of the corpus, as well as the poor quality of the thesaurus used.
 - Further investigations using a better thesaurus should be carried out.

Finally, the comparative studies of the TREC results showed that *Oracle® Text*'s retrieval performance for Chinese IR is comparable to those of the TREC-5 participants. *Oracle® Text* has the best overall average precision, as well as the second best R-precision and Document Level Precision readings for its Manual Run. The overall results of its Automatic Run are also satisfactory, scoring above-average recall, average precision and R-precision readings.

8 Conclusions and Recommendations

The objective of this thesis was to study the issues concerning Chinese IR, and to investigate factors that affect the retrieval effectiveness of a Chinese IR system. Based on the evaluation framework of the TREC-5 Chinese Track and the use of a large corpus of Chinese news articles, a series of experiments were carried out on an *Oracle10g* (beta version) database that was customized to process simplified Chinese characters. Finally, the performance of *Oracle® Text* in Chinese IR was benchmarked through a comparative evaluation made against the retrieval results of the TREC-5 participants. Although the experimental environment was limited, the observations gained could serve as a basis for future investigations.

8.1 Issues in Chinese IR

In the Chinese language, a sentence is written as a string of characters without any word boundary. Hence, the indexing approaches of Chinese IR systems are based on single characters, n-grams or word segmentation techniques. Although each of these approaches produces different types of tokens, and therefore affects IR differently, the observations of the experiments showed that the major factors influencing retrieval performance lie in the ways which queries are formulated – for example, the number of query terms used, the choice of search operators, as well as whether synonyms and all possible transliterations or abbreviations are included in a search.

As discussed in Chapters 2 and 3, besides the two standards of character forms (i.e. simplified Chinese and traditional Chinese) that are used in the modern Chinese language, there exist several regional differences in the Chinese vocabulary, transliteration of non-Chinese proper names, as well as the ways which Chinese characters are encoded for information processing on computer systems.

8.2 Potential of *Oracle® Text* in Supporting Chinese IR

As shown in Chapter 6, the *Oracle* database supports all major Chinese character sets and encoding standards, and its NLS architecture enables simultaneous support of different Chinese client locales.

A full-text index (*context* index) can be created on Chinese texts using either the *Chinese_Lexer* or *Chinese_Vgram_Lexer*. The *Chinese_Lexer*, which uses a hybrid algorithm combining dictionary-based word segmentation and bi-grams indexing, has the advantage of generating real word tokens and a smaller index, as well as producing retrieval results of higher precision. Although stopwords are supported in the *Chinese_Lexer*, it has not been proven that retrieval effectiveness could be influenced through the use of stopwords.

In terms of query formulation, the ACCUM, NEAR and Weight (*) operators are found to be especially effective for Chinese IR. It was believed that retrieval performance could be improved through the use of theme operators. However, no suitable thesaurus was available at the point in time, when the experiments were conducted. Likewise, the knowledge base supplied by *Oracle® Text* was only limited to English and French. Although not evaluated within the scope of this thesis, it is assumed that the WITHIN operator could improve precision in the IR for structured Chinese texts, since this operator allows the searching of keywords within a sentence, a paragraph or a pre-defined section.

8.3 Recommendations for Future Research

It is evident that the ways in which a query is formulated are critical to retrieval performance, and it is important to include in a query, all necessary synonyms in the form of abbreviations of proper nouns or short phrases, as well as to take into account, all possible transliterations of a given non-Chinese proper name.

Although it is believed that the use of a thesaurus could improve query formulation and retrieval performance, no concrete results have been obtained in this thesis to support the hypothesis. Therefore, a further investigation in this area is recommended.

Experiment V has also shown that when no feedback or user interactivity are present in an IR process, the overall retrieval performance is far worse than that of another IR process, where extensive use of synonyms and revision of queries are involved. Hence, features to support user-centered interactive query expansion could be explored.

Like in the IR of all other languages, the trend in Chinese IR is to support natural language searching. It is therefore necessary to study how meaningful keywords could be extracted from queries, and how the regional differences in the Chinese vocabulary could be accommodated in a Chinese knowledge base.

Finally, attention could be drawn to the study of users' search behavior in Chinese IR, and its design implications for the user interface.

Appendix A: Topic Statements of TREC-5 Chinese Track

Source: TREC Data – Non-English Test Questions (Topics) Files List.
(URL: http://trec.nist.gov/data/topics_noneng/index.html)

- Number: CH1
Title: 美国决定将中国大陆的人权状况与其是否给予中共最惠国待遇分离。
U.S. to separate the most-favored-nation status from human rights issue in China.
Description: 最惠国待遇, 中国, 人权, 经济制裁, 分离, 脱钩
most-favored nation status, human rights in China, economic sanctions, separate, untie
Narrative: 相关文件必须提到美国为何将最惠国待遇与人权分离; 相关文件也必须提到中共为什么反对美国将人权与最惠国待遇相提并论。
A relevant document should describe why the U.S. separates most-favored nation status from human rights. A relevant document should also mention why China opposes U.S. attempts to tie human rights to most-favored-nation status.
- Number: CH2
Title: 中共对于中国统一的立场
Communist China's position on reunification
Description: 中国, 一国两制, 台湾, 和平统一, 经贸合作, 两岸关系, 科技、文化交流
China, one-nation-two-systems, Taiwan, peaceful reunification, economic and trade cooperation, cross-strait relationship, science and technology exchanges
Narrative: 相关文件必须提到中共如何经由实现一国两制来达到台湾与大陆统一的目的。如果文件只是外国政府重申支持中共对台湾拥有主权或提到中共与其他国家之经贸、科技、文化交流, 则为不相关文件。
A relevant document should describe how China wishes to reach reunification through the implementation of "one-nation-two-systems." If a document merely states a foreign nation's support of China's sovereignty over Taiwan or discusses trade cooperation as well as cultural and technical exchanges between China and a country other than Taiwan, then the document is irrelevant.
- Number: CH3
Title: 中共核电站之营运情况
The operational condition of nuclear power plants in China
Description: 核电站, 大亚湾, 秦山, 安全
nuclear power plant, Daya Bay (nuclear power plant), Qinshan (nuclear power plant), safety
Narrative: 相关文件必须提到中国目前投产的核电站的安全营运情况。任何有关安全之规则或法令, 安全措施之执行, 意外事故报告之文件皆属相关文件。
A relevant document should contain information on the current safety practices in China's nuclear power plants. Any article on safety regulations, accident reports and safety practices are relevant.
- Number: CH4
Title: 中国大陆新发现的油田
The newly discovered oil fields in China
Description: 油田, 天然气, 油气, 储量, 油质
oil field, natural gas, oil and gas, oil reserves, oil quality
Narrative: 相关文件必须提到中国大陆近几发现的油田的储量, 各油田的特点, 以及中国开发油田的计划。
A relevant document should contain information on the oil reserves in the newly discovered oil fields in Mainland China, any concrete description of specific oil fields, or China's plan to develop these fields.
- Number: CH5
Title: 中国有关知识产权的立法与政策以及执法情况
Regulations and Enforcement of Intellectual Property Rights in China
Description: 知识产权法, 商标法, 著作权法, 专利法
intellectual property rights, trade mark, copyright, patent
Narrative: 相关文件必须提到中国有关保护知识产权的法律。非相关文件包括将中国违反知识产权作为

对中国贸易制裁之依据或中国以知识产权作为经济改革的项目。

A relevant document should describe laws established in China to protect intellectual property rights. If a document contains information such as: China's violation of intellectual property rights as the basis for imposing trade sanctions against China; or, China taking up intellectual property rights as part of its economic reform, then the document is irrelevant.

- Number: CH6
 Title: 国际社会对中共加入世界贸易组织所给予之支持
 International Support of China's Membership in the WTO
 Description: 世界贸易组织, 关贸总协, 市场准入, 世界贸易体系, 多边贸易, 成员(国)
 World Trade Organization (WTO), GATT, market access, world trade structure, multilateral trade, member nation
 Narrative: 相关文件必须提到某一国家或某些国家对中国加入世界贸易组织所给予之支持。
 A relevant document should indicate support given by specific nation(s) for China's membership in WTO.
- Number: CH7
 Title: 中国大陆与台湾对南海诸岛的立场
 Claims made by both PRC and Taiwan over islands in the South China Sea
 Description: 南沙(群岛), 东沙(群岛), 西沙(群岛), 中国, 台湾, 主权
 The Spratly Islands, the Dongsha Islands, the Xisha Islands, China, Taiwan, sovereignty
 Narrative: 相关文件应包括下列信息: (1)为何南沙群岛成为中国、菲律宾、越南、印尼等国冲突的所在地; (2)南海有那些天然资源; (3)中国大陆与台湾对南海诸岛之主权立场为何; 以及(4)东盟国家对解决南沙群岛与南海争端有什么建议。
 A relevant document should include the following information: (1) why the Spratly Islands became the disputed area among China, the Philippines, Vietnam, and Indonesia; or (2) what are the natural resources found in the South China Sea; or/and (3) what are the sovereign rights claimed by the PRC and Taiwan; or/and (4) what are the suggestions proposed by the ASEAN to solve the territorial dispute over the Spratly Islands and South China Sea.
- Number: CH8
 Title: 地震在日本造成的损害与伤亡数据
 Numeric Indicators of Earthquake Severity in Japan
 Description: 日本, 地震, 损失, 死亡, 级, 受伤, 芮氏地震仪
 Japan, earthquake, damage, death, injury, Richter scale
 Narrative: 相关文件应包括地震的级数以及所造成的实际损害与伤亡数字, 诸如地震在芮氏地震仪上的级数, 死亡与受伤人数, 以及以金钱为单位的财产损失数目。
 A relevant document should contain numeric indicators such as the magnitude of the earthquake, number of deaths or injuries, or property damage.
- Number: CH9
 Title: 中国毒品问题
 Drug Problems in China
 Description: 毒品, 可卡因, 大麻, 海洛因, 吨, 公吨, 吸食毒品, 毒品买卖
 narcotics, cocaine, heroin, marijuana, ton(s), kilogram(s), drugs use, drugs sale
 Narrative: 相关文件应包括目前毒品在中国所造成的危害, 中国打击非法买卖毒品的措施, 是否有戒毒设施, 以及中国是否与国际执法组织合作来遏制国际毒贩的走私活动。
 A relevant document should contain information on drug problems in China, how the government cracks down on illegal drug activities, what types of drug rehabilitation program exist in China, and how the Chinese government cooperates with international organizations to stop the spread of drug trafficking.
- Number: CH10
 Title: 新疆的边境贸易
 Border Trade in Xinjiang
 Description: 新疆, 维吾尔, 边境贸易, 边贸, 市场
 Xinjiang, Uigur, border trade, market
 Narrative: 相关文件必须包括中国新疆与其邻近国家的贸易关系, 此关系包括中国与前苏联共和国之间所签署的贸易条约以及彼此间的外贸投资。如果文件只论及中国如何建设发展新疆, 则属非相关文件。
 A relevant document should contain information on the trading relationship between Xinjiang, China and its neighboring nations, including treaties signed by China and former Soviet Republics that are bordering China and foreign investment. If a document contains information on how China develops Xinjiang, it is not relevant.

- Number: CH11
 Title: 联合国驻波斯尼亚维和部队
 UN Peace-keeping Force in Bosnia
 Description: 波斯尼亚, 前南斯拉夫, 巴尔干, 联合国, 北约, 武器禁运, 维和, 维持和平
 Bosnia, Former Yugoslavia, Balkan, U.N., NATO, Muslim, weapon sanction, peace-keeping
 Narrative: 相关文件必须包括联合国和平部队如何在战火蹂躏的波斯尼亚进行维持和平的任务。
 A relevant document should contain information on how UN peace-keeping troops carry out their mission in the war-torn Bosnia.
- Number: CH12
 Title: 世界妇女大会
 World Conference on Women
 Description: 联合国, 世界, 妇女大会, 妇女问题, 妇女地位
 UN, world, women's conference, women's issues, women's status
 Narrative: 相关文件必须是关于第四届世界妇女大会中讨论的妇女问题,特别是经由教育和立法来改进妇女的社会地位和经济情况的措施。
 A relevant document should contain information on the 4th World Conference on Women, especially on ways to improve women's social status and economic situations through education and legislation.
- Number: CH13
 Title: 中国争取举办西元 2000 年奥运
 China Bids for 2000 Olympic Games
 Description: 中国, 经济实力, 奥运, 世界运动大会, 奥林匹克, 筹备工作
 China, economic strength, Olympic games, preparatory work
 Narrative: 相关文件必须包括中国如何争取举办西元 2000 年奥运,中国所持的理由为何。中国选手在奥运会中的表现属于不相关文件。
 A relevant document should contain information on how China bids for the 2000 Olympic Games, China's reasons for sponsoring the 2000 Olympic games.
- Number: CH14
 Title: 中国的爱滋病例
 Cases of AIDS in China
 Description: 中国, 云南, 爱滋病, HIV, 高危险群患者, 注射器, 病毒
 China, Yunnan, AIDS, HIV, high risk group, syringe, virus
 Narrative: 相关文件应当包括中国那些地区的爱滋病例最多, 爱滋病毒在中国是如何传播的, 以及中国政府如何监测爱滋病并控制它的传染。
 A relevant document should contain information on the areas in China that have the highest AIDS cases, how the AIDS virus was transmitted, and how the Chinese government combats AIDS problem.
- Number: CH15
 Title: 联合国维和部队如何帮助海地恢复民主制度
 The UN peace-keeping troops help Haiti return to democracy
 Description: 海地, 联合国, 美国, 多国部队, 维和部队, 民主
 Haiti, UN, U.S., multination-troops, peace-keeping troops, democracy
 Narrative: 相关文件必须提到美国如何帮助海地民主政府重建海地; 联合国安理会对海地问题之决议, 以及拉美国对联合国决议之反应。不相关文件则为海地仅为新闻或电视广播提要, 或新闻分析中提及海地但新闻主题不在海地。
 A relevant document should contain information on the U.S. efforts to help Haiti resume its democracy, UN resolutions on Haiti, and the Latin-American nations' reactions to the UN resolutions.
- Number: CH16
 Title: 联合国对伊拉克经济制裁的辩论
 The Debate of UN Sanctions Against Iraq
 Description: 联合国, 伊拉克, 经济制裁
 UN, Iraq, economic sanction
 Narrative: 相关文件应提到联合国为何对伊拉克实施经济制裁; 经济制裁对伊拉克的影响; 联合国对何时解除此经济制裁的辩论; 以及伊拉克对经济制裁的反应。不相关文件为法国为了在伊拉克设代表处而减少其对伊之经济制裁; 中国对联合国在中东维和行动的评论; 伊拉克与伊朗关

系正常化中批评联合国之制裁；或联合国对伊拉克之经济制裁仅为新闻提要而未详细报道。 A relevant document should contain information on why the UN carries out economic sanctions against Iraq; the impact of the economic sanctions on Iraq; the UN debate on when to lift the sanctions; Iraq's reaction to the sanctions. An irrelevant document is such that it only mentions the UN sanctions against Iraq but does not give any details on the discussions, impact, and Iraq's reaction about the sanctions. Non-relevant documents include summaries without any details like the French government's setting up a representative office in Iraq thus reducing its economic sanctions toward Iraq; Iran's criticizing of the UN sanctions when seeking diplomatic relations with Iraq, or UN sanctions against Iraq.

- Number: CH21
 Title: 香港总督彭定康在香港回归中国一事上所扮演的角色
 The Role of the Governor of Hong Kong in the Reunification with the PRC
 Description: 香港问题，特别行政区，彭定康，计划，建议
 Hong Kong issue, special administrative zone, Peng DingKang, plan, proposal
 Narrative: 相关文件：应包括香港总督彭定康在香港问题上所扮演的角色，包括所有彭定康发表过的声明，彭定康到中国访问与中国政府官员的谈话，以及中国政府对彭定康提出的有关香港立法改革的批评等。不相关文件：任何非来自香港，英国，或中国的有关彭定康的评论皆属非相关文件。
 A relevant document presents information on the role of the Governor of Hong Kong, Peng DingKang, in the reunification of China. Issues include any of the Governor's announcements, his official visits to China and meetings with Chinese officials, PRC criticism of Peng's legislative plans or proposals, etc. Non-relevant documents discuss any reactions to the Governor's actions or his politics in Hong Kong reunification from sources other than Hong Kong, UK, or the PRC.
- Number: CH22
 Title: 世界各地感染疟疾的情况
 The Spread of Malaria Infection in Various Parts of the World
 Description: 疟疾，死亡人数，感染病例
 malaria, number of deaths, number of infections
 Narrative: 相关文件应包括有关世界各地感染疟疾的情况，包括病例统计与死亡人数。凡属讨论与传染性疾病预防有关的卫生政策或预防疟疾之疫苗接种而未提及感染或死亡人数的资料则为非相关文件。
 A relevant document presents numeric information about malaria infection or death rate at a national or international level. Non-relevant documents discuss health policies related to communicable diseases or vaccination against malaria without numeric information.
- Number: CH23
 Title: 苏联在海湾战争中如何担任调停的角色
 Soviet Union's Mediation Role in the Gulf War
 Description: 苏联，海湾战争，和平建议，伊拉克
 Soviet Union, Gulf War, peace proposal, Iraq
 Narrative: 相关文件应提及苏联在海湾战争中如何担任调停的角色，包括与伊拉克之间的沟通，苏联在联合国安理会中提出的停火协议以及要求多国部队从伊拉克撤出的和平建议。
 A relevant document discusses the Soviet Union's mediation in the Gulf War, including communication with Iraq, cease-fire resolution to the UN Security Council and their peace proposal for withdrawal of multi-national troops, etc.

Appendix B: Index Description

The followings are description of the two indices used in the experiments described in Chapter 7.

B.1 *ch_index*

```

=====
INDEX DESCRIPTION
=====

index name:                "SCOTT"."CH_INDEX"
index id:                   1414
index type:                 context
base table:                 "SCOTT"."TREC"
primary key column:
text column:               DOCS
text column type:         CLOB
language column:
format column:
charset column:

=====
INDEX OBJECTS
=====

datastore:                 DIRECT_DATASTORE

filter:                    NULL_FILTER

section group:             BASIC_SECTION_GROUP
  zone section:            TEXT
    section tag:          TEXT
  zone section:            HEADLINE
    section tag:          HL
  zone section:            HEADLINE
    section tag:          HEADLINE
  zone section:            PARA
    section tag:          P
  zone section:            LINE
    section tag:          S
  zone section:            DOC
    section tag:          DOC
  field section:          DOCNO
    section tag:          DOCNO
    visible:              N
    field id:             17
  field section:          DATE
    section tag:          DATE
    visible:              N
    field id:             18
  field section:          DOCID
    section tag:          DOCID
    visible:              N
    field id:             16

lexer:                     CHINESE_LEXER

wordlist:                  BASIC_WORDLIST
  stemmer:                NULL
  fuzzy_match:            CHINESE_VGRAM

```

stoplist:	BASIC_STOPLIST
stop_word:	一再
stop_word:	一同
stop_word:	一得
stop_word:	下去
stop_word:	两个
stop_word:	为了
stop_word:	之一
stop_word:	也是
stop_word:	什么
stop_word:	从而
stop_word:	他们
stop_word:	以上
stop_word:	以及
stop_word:	但是
stop_word:	共同
stop_word:	具有
stop_word:	即可
stop_word:	却是
stop_word:	可以
stop_word:	可能
stop_word:	各方面
stop_word:	各种
stop_word:	哪里
stop_word:	回到
stop_word:	因为
stop_word:	因此
stop_word:	基于
stop_word:	基本上
stop_word:	处在
stop_word:	大量
stop_word:	如何
stop_word:	如果
stop_word:	存在着
stop_word:	它的
stop_word:	实在
stop_word:	对于
stop_word:	就是
stop_word:	尽管
stop_word:	已经
stop_word:	带来
stop_word:	带着
stop_word:	并非
stop_word:	当时
stop_word:	很少
stop_word:	很有
stop_word:	得到
stop_word:	必将
stop_word:	必须
stop_word:	成为
stop_word:	我们
stop_word:	或者
stop_word:	所有
stop_word:	所需
stop_word:	新的
stop_word:	方面
stop_word:	来自
stop_word:	此项
stop_word:	没有

```

stop_word:          现在
stop_word:          由于
stop_word:          目前
stop_word:          相当
stop_word:          而言
stop_word:          获得了
stop_word:          许多
stop_word:          超过
stop_word:          较大
stop_word:          达到
stop_word:          还将
stop_word:          还是
stop_word:          还有
stop_word:          这个
stop_word:          那里
stop_word:          都是
stop_word:          除了

storage:            BASIC_STORAGE
  r_table_clause:
lob (data) store as (cache)
  i_index_clause:   compress 2

```

B.2 zh_index

```

=====
INDEX DESCRIPTION
=====

```

```

index name:         "SCOTT"."ZH_INDEX"
index id:           2792
index type:         context
base table:         "SCOTT"."TREC"
primary key column:
text column:        DOCS
text column type:   CLOB
language column:
format column:
charset column:

```

```

=====
INDEX OBJECTS
=====

```

```

datastore:          DIRECT_DATASTORE

filter:              NULL_FILTER

section group:      BASIC_SECTION_GROUP
  zone section:     TEXT
    section tag:    TEXT
  zone section:     HEADLINE
    section tag:    HL
  zone section:     HEADLINE
    section tag:    HEADLINE
  zone section:     PARA
    section tag:    P
  zone section:     LINE
    section tag:    S

```


zone section:	DOC
section tag:	DOC
field section:	DOCNO
section tag:	DOCNO
visible:	N
field id:	17
field section:	DATE
section tag:	DATE
visible:	N
field id:	18
field section:	DOCID
section tag:	DOCID
visible:	N
field id:	16
lexer:	CHINESE_VGRAM_LEXER
wordlist:	BASIC_WORDLIST
stemmer:	NULL
fuzzy_match:	CHINESE_VGRAM
stoplist:	BASIC_STOPLIST
stop_word:	一再
stop_word:	一同
stop_word:	一得
stop_word:	下去
stop_word:	两个
stop_word:	为了
stop_word:	之一
stop_word:	也是
stop_word:	什么
stop_word:	从而
stop_word:	他们
stop_word:	以上
stop_word:	以及
stop_word:	但是
stop_word:	共同
stop_word:	具有
stop_word:	即可
stop_word:	却是
stop_word:	可以
stop_word:	可能
stop_word:	各方面
stop_word:	各种
stop_word:	哪里
stop_word:	回到
stop_word:	因为
stop_word:	因此
stop_word:	基于
stop_word:	基本上
stop_word:	处在
stop_word:	大量
stop_word:	如何
stop_word:	如果
stop_word:	存在着
stop_word:	它的
stop_word:	实在
stop_word:	对于
stop_word:	就是
stop_word:	尽管
stop_word:	已经
stop_word:	带来
stop_word:	带着

stop_word:	并非
stop_word:	当时
stop_word:	很少
stop_word:	很有
stop_word:	得到
stop_word:	必将
stop_word:	必须
stop_word:	成为
stop_word:	我们
stop_word:	或者
stop_word:	所有
stop_word:	所需
stop_word:	新的
stop_word:	方面
stop_word:	来自
stop_word:	此项
stop_word:	没有
stop_word:	现在
stop_word:	由于
stop_word:	目前
stop_word:	相当
stop_word:	而言
stop_word:	获得了
stop_word:	许多
stop_word:	超过
stop_word:	较大
stop_word:	达到
stop_word:	还将
stop_word:	还是
stop_word:	还有
stop_word:	这个
stop_word:	那里
stop_word:	都是
stop_word:	除了
storage:	BASIC_STORAGE
r_table_clause:	
lob (data) store as (cache)	
i_index_clause:	compress 2

Appendix C: Search Statements (Queries)

Below are the search statements for the Automatic and Manual runs in Experiment I, II and III:

C.1 Experiment I & III

Topic	Automatic Run	Manual Run
CH1	最惠国待遇, 中国, 人权, 经济制裁, 分离, 脱钩	美国 AND 中国 AND 最惠国待遇
CH2	中国, 一国两制, 台湾, 和平统一, 经贸合作, 两岸关系, 科技, 文化, 交流	(台湾 海峡两岸) AND (两岸关系 两岸经贸 和平统一 一国两制 主权)
CH3	核电站, 大亚湾, 秦山, 安全	((核电站 核电厂) AND (大亚湾 秦山)) AND 安全
CH4	油田, 天然气, 油气, 储量, 油质	(油田 天然气 油气 石油) AND (储量 藏量 油质) AND 中国
CH5	知识产权法, 商标法, 著作权法, 专利法	(知识产权 商标法 著作权法 专利法) AND (中国 中共)
CH6	世界贸易组织, 关贸总协, 市场准入, 世界贸易体系, 多边贸易, 成员	(世界贸易组织 关贸总协 市场准入 缔约国) AND 中国 AND (支持 ACCUM (成员 美国 德国 法国 欧洲))
CH7	南沙, 东沙, 西沙, 中国, 台湾, 主权	(南沙 东沙 西沙 南海 海南岛) AND (主权 资源 油田) AND (中国 中共 台湾 东盟国 菲律宾 越南 印尼)
CH8	日本, 地震, 损失, 死亡, 级, 受伤, 芮氏地震仪	日本 AND 地震
CH9	毒品, 可卡因, 大麻, 海洛因, 吨, 公吨, 吸食毒品, 毒品买卖	(毒品 可卡因 大麻 海洛因 毒品买卖 吸食毒品 戒毒 毒贩 走私活动) AND (中国 中共)
CH10	新疆, 维吾尔, 边境贸易, 边贸, 市场	(新疆 维吾尔) AND ((边境贸易 边贸) ACCUM 苏联)
CH11	波斯尼亚, 前南斯拉夫, 巴尔干, 联合国, 北约, 武器禁运, 维和, 维持和平	(波斯尼亚 前南斯拉夫 巴尔干) AND (和平部队 维和 武器禁运)
CH12	联合国, 世界, 妇女大会, 妇女问题, 妇女地位	(世界妇女大会 世妇会) AND (妇女问题, 妇女地位, 社会地位, 提高地位, 经济情况, 妇女解放, 教育, 立法)
CH13	中国, 经济实力, 奥运, 世界运动大会, 奥林匹克, 筹备工作	(中国 北京) AND (奥运 奥林匹克) AND (举办 举办权)
CH14	中国, 云南, 爱滋病, HIV, 高危险群患者, 注射器, 病毒	(中国 云南 昆明) AND (爱滋病 艾滋病)
CH15	海地, 联合国, 美国, 多国部队, 维和部队, 民主	海地 NOT 上海 NOT 民主党 NOT 海地区 AND (民主*3, 制裁*3, 石油禁运*3, 联合国, 美国)
CH16	联合国, 伊拉克, 经济制裁	联合国 AND 伊拉克 AND (经济制裁 制裁 石油禁运)
CH21	香港问题, 特别行政区, 彭定康, 计划, 建议	(香港问题 特别行政区 香港立法 政改方案 回归 过渡时期 基本法 一国两制) AND (彭定康 香港总督 港督)
CH22	疟疾, 死亡人数, 感染病例	疟疾
CH23	苏联, 海湾战争, 和平建议, 伊拉克	苏联 AND (海湾战争, 停火协议, 撤军, 和平) AND 伊拉克

C.2 Experiment II

Topic	Automatic Run (control set)	Automatic Run (with Weighted Terms)
CH1	最惠国待遇, 中国, 人权, 经济制裁, 分离, 脱钩	最惠国待遇*4, 中国*2, 人权, 经济制裁, 分离, 脱钩
CH2	中国, 一国两制, 台湾, 和平统一, 经贸合作, 两岸关系, 科技, 文化, 交流	中国*2, 一国两制*2, 台湾*3, 和平统一, 经贸合作, 两岸关系, 科技*2, 文化*2, 交流*2
CH3	核电站, 大亚湾, 秦山, 安全	核电站*3, 大亚湾, 秦山, 安全
CH4	油田, 天然气, 油气, 储量, 油质	油田*4, 天然气, 油气, 储量*2, 油质
CH5	知识产权法, 商标法, 著作权法, 专利法	知识产权法, 商标法, 著作权法, 专利法
CH6	世界贸易组织, 关贸总协, 市场准入, 世界贸易体系, 多边贸易, 成员	世界贸易组织*3, 关贸总协, 市场准入, 世界贸易体系, 多边贸易, 成员
CH7	南沙, 东沙, 西沙, 中国, 台湾, 主权	南沙*3, 东沙, 西沙, 中国*4, 台湾*2, 主权*2
CH8	日本, 地震, 损失, 死亡, 级, 受伤, 芮氏地震仪	日本*2, 地震*4, 损失, 死亡*2, 级*2, 受伤*2, 芮氏地震仪*2
CH9	毒品, 可卡因, 大麻, 海洛因, 吨, 公吨, 吸食毒品, 毒品买卖	毒品*4, 可卡因, 大麻, 海洛因, 吨, 公吨, 吸食毒品, 毒品买卖
CH10	新疆, 维吾尔, 边境贸易, 边贸, 市场	新疆*4, 维吾尔, 边境贸易*2, 边贸, 市场
CH11	波斯尼亚, 前南斯拉夫, 巴尔干, 联合国, 北约, 武器禁运, 维和, 维持和平	波斯尼亚*3, 前南斯拉夫, 巴尔干, 联合国*3, 北约, 武器禁运, 维和, 维持和平*2
CH12	联合国, 世界, 妇女大会, 妇女问题, 妇女地位	联合国, 世界*3, 妇女大会*3, 妇女问题*2, 妇女地位
CH13	中国, 经济实力, 奥运, 世界运动大会, 奥林匹克, 筹备工作	中国*4, 经济实力, 奥运*4, 世界运动大会, 奥林匹克, 筹备工作
CH14	中国, 云南, 爱滋病, HIV, 高危险群患者, 注射器, 病毒	中国*5, 云南, 爱滋病*5, HIV, 高危险群患者, 注射器, 病毒*2
CH15	海地, 联合国, 美国, 多国部队, 维和部队, 民主	海地*7, 联合国*4, 美国*2, 多国部队, 维和部队*2, 民主*3
CH16	联合国, 伊拉克, 经济制裁	联合国*7, 伊拉克*8, 经济制裁*8
CH21	香港问题, 特别行政区, 彭定康, 计划, 建议	香港问题*2, 特别行政区, 彭定康*7, 计划, 建议
CH22	疟疾, 死亡人数, 感染病例	疟疾*3, 死亡人数*2, 感染病例
CH23	苏联, 海湾战争, 和平建议, 伊拉克	苏联*4, 海湾战争*3, 和平建议*2, 伊拉克*3

C.3 Experiment IV

Topic	Original Manual Run in Experiment I	Manual Run with Thesaurus Terms
CH3	(核电站 核电厂) AND(大亚湾 秦山) AND 安全	SYN(核能) AND(大亚湾 秦山)
CH5	(知识产权 商标法 著作权法 专利法) AND (中国 中共)	(SYN(知识产权) BT(知识产权)) AND (中国 中共)
CH13	(中国 北京) AND (奥运 奥林匹克) AND (举办 举办权)	(中国 北京) AND SYN(奥运会) AND (举办 举办权)
CH23	苏联 AND (海湾战争, 停火协议, 撤军, 和平) AND 伊拉克	苏联 AND (SYN(波斯湾战争及其影响) BT(波斯湾战争及其影响)) AND 伊拉克

The relationships of the selected thesaurus terms are as follows:

核能

UF NUCLEAR ENERGY
 NT 反核子示威
 NT 反核子武器示威
 NT 核废料
 NT 核辐射
 NT 核燃料
 NT 核子厂
 NT 核子防空壕
 NT 核子试炸
 NT 核子武器
 NT 核子意外事件

知识产权

SYN 专利权
 BT 版权

奥运会

SYN 奥林匹克运动会
 BT 运动会

波斯湾战争及其影响

SYN 伊拉克侵略科威特
 UF GULF WAR
 BT 战争
 BT 中东

C.4 Experiment V

<i>Topic</i>	<i>User</i>	<i>Query</i>
CH1	User 1	美国, 中国, (人权状况 AND 最惠国)*2
	User 2	美国 AND 中国 AND (人权状况 AND 最惠国)
	User 3	中国 AND 最惠国
	User 4	美国, 中国, 人权*2
	User 5	美国; 中国人权
	User 6	中国 AND 人权 AND 最惠国
CH2	User 1	中国 AND 一国两制 AND 统一
	User 2	(中共 OR 中国) AND 统一立场
	User 3	中国 AND 统一
	User 4	中共 AND 统一
	User 5	中国; 统一
	User 6	中国 AND 统一 AND 一国两制 NOT 台湾主权
CH3	User 1	核电站 AND 中国
	User 2	核电站 AND 营运
	User 3	中共 AND 核电站
	User 4	中共 AND 核电站
	User 5	中共; 核电站
	User 6	中共 AND 核电站 AND 安全
CH4	User 1	中国 AND 油田
	User 2	中国 AND (发现 AND 油田)
	User 3	中国 AND 油田
	User 4	中国 AND 油田
	User 5	中国; 新油田
	User 6	中国 AND 新油田
CH5	User 1	保护, (中国 AND 知识产权立法)
	User 2	中国 AND 知识产权立法 AND 政策
	User 3	中国 AND 知识产权 AND 立法 AND 执法
	User 4	中国 AND 知识产权
	User 5	中国; 知识产权; 立法; 执法
	User 6	中国 AND 知识产权 AND 保护 NOT (贸易 OR 经济改革)
CH6	User 1	中国, 世界贸易组织, 加入*2
	User 2	中共 AND 世界贸易组织 AND 支持
	User 3	中国 AND 世界贸易组织 AND 加入
	User 4	中国 AND 世界贸易组织*2
	User 5	中国; 世界贸易组织
	User 6	中国 AND(世界贸易组织 OR 世贸) AND 加入

CH7	User 1	中国, 台湾, 冲突*2 AND 南海*2
	User 2	南海诸岛 AND (中国 AND 台湾)
	User 3	(大陆 OR 中国 OR 台湾) AND 南海群岛
	User 4	大陆, 台湾, 南海群岛
	User 5	中国; 南海群岛
	User 6	中国 AND 台湾 AND (南海 OR 南沙群岛) AND (冲突 OR 天然资源 OR 主权 OR 东盟)
CH8	User 1	(日本 AND 地震), 数据, 数字
	User 2	日本 AND 地震 AND (损害 OR 伤亡)
	User 3	日本 AND 地震 AND 数据
	User 4	日本, 地震, 数据
	User 5	日本地震
	User 6	日本 AND 地震 AND (损 OR 伤亡)
CH9	User 1	毒品问题 AND 中共
	User 2	中国毒品问题
	User 3	中国 AND 毒品
	User 4	中国 AND 毒品
	User 5	中国; 毒品
	User 6	中国 AND 毒品
CH10	User 1	新疆 AND 边境贸易*2 NOT 建设发展
	User 2	新疆 AND 边境贸易
	User 3	新疆 AND 贸易
	User 4	新疆 AND 贸易
	User 5	新疆; 边境贸易
	User 6	新疆 AND 贸易 AND (贸易条约 OR 外贸) NOT (建设 OR 发展)
CH11	User 1	(波斯尼亚 AND 联合国), 和平, 部队
	User 2	波斯尼亚 AND (联合国和平部队 OR 和平任务)
	User 3	波斯尼亚 AND 联合国
	User 4	波斯尼亚 AND 维和部队
	User 5	波斯尼亚; 维和部队; 联合国
	User 6	波斯尼亚 AND 联合国 AND (维和部队 OR 维持和平)
CH12	User 1	第四届世界妇女大会 AND 妇女
	User 2	世界妇女大会 AND (教育 OR 妇女的社会地位)
	User 3	世界妇女大会 AND 第四
	User 4	世界妇女大会 AND 第四届
	User 5	第四届世界妇女大会
	User 6	第四届 AND 世界妇女大会
CH13	User 1	奥运*2 AND 中国
	User 2	奥运 AND 2000 AND 中国 AND 理由 NOT 中国选手表现
	User 3	中国, 奥运, 2000
	User 4	中国, 奥运*2, 2000 年
	User 5	中国申办 2000 年奥运会
	User 6	中国 AND 2000 AND 奥运 NOT 选手

CH14	User 1	中国 AND 爱滋病
	User 2	中国 AND 爱滋病
	User 3	中国 AND 爱滋病
	User 4	中国 AND 爱滋病
	User 5	中国; 艾滋病
	User 6	中国 AND (爱滋病 OR AIDS)
CH15	User 1	(海地*2 AND 联合国), 重建
	User 2	美国 AND (海地民主政府 OR 重建海地)
	User 3	海地 AND 联合国 AND 民主
	User 4	海地 AND (维和部队 OR 民主制度)
	User 5	海地; 联合国; 维和部队
	User 6	联合国 AND (维和部队 OR 安理会) AND 海地 AND 民主
CH16	User 1	(经济制裁, 伊拉克*2, 联合国) NOT 法国 NOT 伊朗
	User 2	联合国 AND 伊拉克 AND 经济制裁 NOT (法国 AND 中国)
	User 3	经济制裁 AND 伊拉克
	User 4	经济制裁*2, 伊拉克, 联合国
	User 5	经济制裁; 伊拉克; 联合国
	User 6	经济制裁 AND 伊拉克 AND 联合国
CH21	User 1	香港回归, 彭定康, 角色
	User 2	香港总督 OR 彭定康) AND (声明 OR 中国访问) AND (中国政府 OR 香港立法改革)
	User 3	香港 AND 总督 AND 回归
	User 4	香港总督 AND 回归
	User 5	香港回归; 彭定康
	User 6	彭定康 AND (香港 OR 英国 OR 中国) AND 回归
CH22	User 1	疟疾, 病例统计, 死亡人数
	User 2	疟疾 AND (病例统计 OR 死亡人数)
	User 3	疟疾 AND 传染
	User 4	疟疾
	User 5	疟疾
	User 6	疟疾 AND 感染 AND (死亡人数 OR 统计) NOT (卫生政策 OR 预防)
CH23	User 1	伊拉克 AND 苏联 AND (停火协议 OR 调停)
	User 2	伊拉克 AND 苏联 AND (停火协议 OR 调停)
	User 3	苏联 AND 海湾战争
	User 4	苏联, 海湾战争, 调停
	User 5	苏联; 海湾战争
	User 6	苏联 AND 海湾战争 AND 角色

Appendix D: Search Results

Here are detailed search results of all the experiments.

D.1 Experiment I

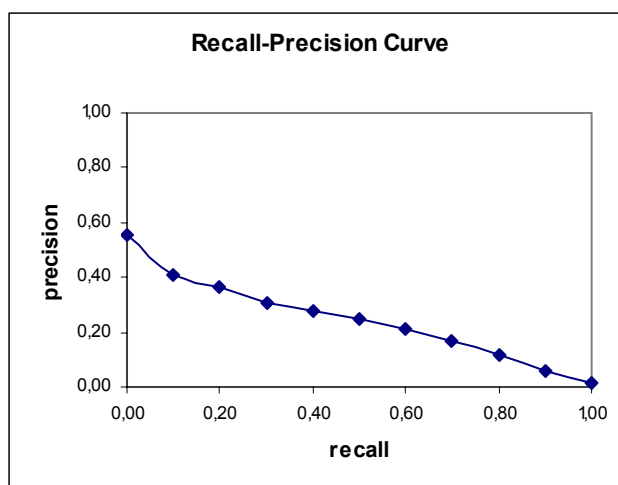
(a) Result Set: ch_auto

Summary of Statistics

Indexing: Chinese Lexer
 Retrieval: Auto (TREC-defined descriptors connected by ACCUM operator)
 Nr. of Topics: 19
 Total Docs retrieved: 17502
 Total relevant Docs: 1399
 Relevant Docs retrieved: 1027

Recall Level Precision Averages

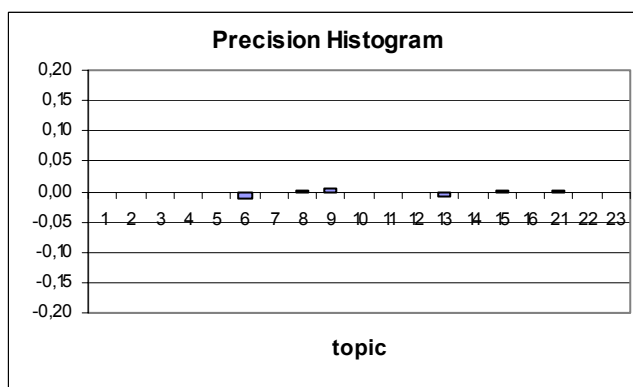
Recall	Precision
0.00	0.5578
0.10	0.4053
0.20	0.3678
0.30	0.3076
0.40	0.2786
0.50	0.2453
0.60	0.2104
0.70	0.1674
0.80	0.1204
0.90	0.0617
1.00	0.0113



Average precision over all relevant docs: 0.2863 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.3789
At 10 docs	0.3842
At 15 docs	0.3649
At 20 docs	0.3579
At 30 docs	0.4772
At 100 docs	0.2311
At 200 docs	0.1634
At 500 docs	0.0927
At 1000 docs	0.0971



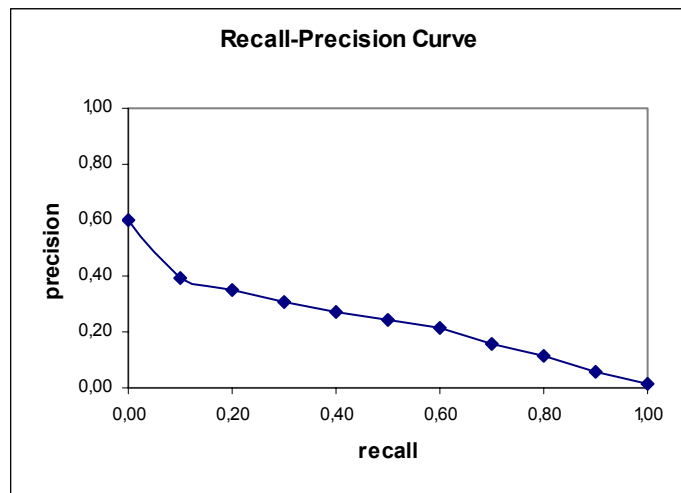
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.2886

(b) Result Set: zh_auto**Summary of Statistics**

Indexing: Chinese Vgram Lexer
 Retrieval: Auto (TREC-defined descriptors connected by ACCUM operator)
 Nr. of Topics: 19
 Total Docs retrieved: 17503
 Total relevant Docs: 1399
 Relevant Docs retrieved: 1055

Recall Level Precision Averages

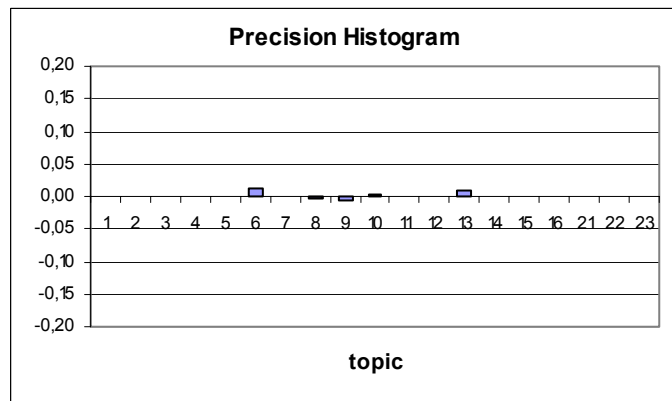
Recall	Precision
0.00	0.5975
0.10	0.3953
0.20	0.3498
0.30	0.3090
0.40	0.2721
0.50	0.2398
0.60	0.2137
0.70	0.1560
0.80	0.1145
0.90	0.0588
1.00	0.0166



Average precision over all relevant docs: 0.2728 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.3789
At 10 docs	0.3632
At 15 docs	0.3439
At 20 docs	0.3316
At 30 docs	0.4623
At 100 docs	0.2089
At 200 docs	0.1661
At 500 docs	0.0938
At 1000 docs	0.0555



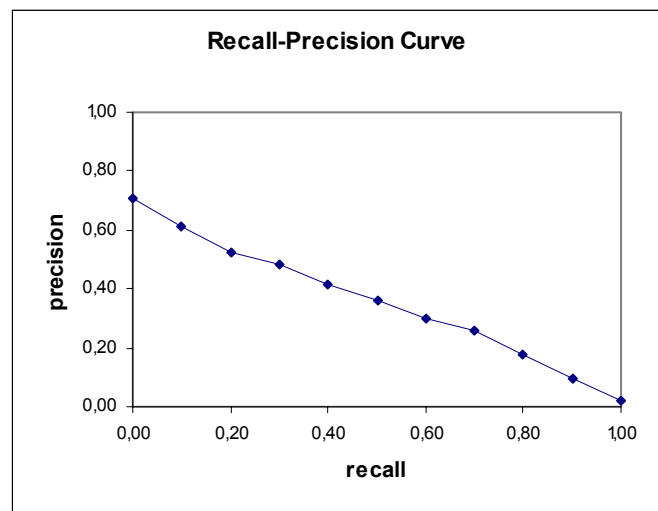
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.2888

(c) Result Set: ch_manual**Summary of Statistics**

Indexing: Chinese Lexer
 Retrieval: Manual (Boolean and ACCUM operators + user defined search terms)
 Nr. of Topics: 19
 Total Docs retrieved: 4280
 Total relevant Docs: 1399
 Relevant Docs retrieved: 951

Recall Level Precision Averages

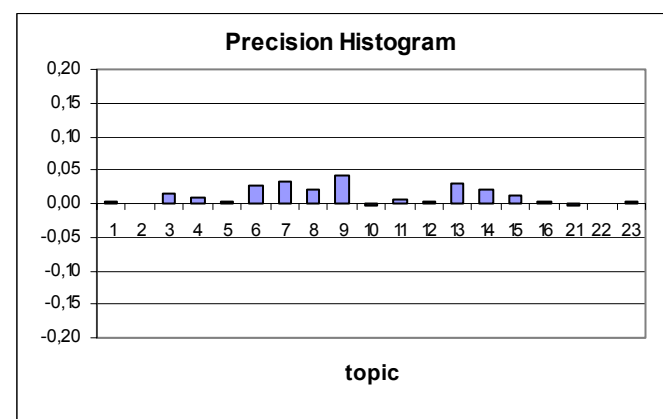
Recall	Precision
0.00	0.7108
0.10	0.6094
0.20	0.5226
0.30	0.4825
0.40	0.4160
0.50	0.3609
0.60	0.3006
0.70	0.2602
0.80	0.1753
0.90	0.0923
1.00	0.0208



Average precision over all relevant docs: 0.4706 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.6526
At 10 docs	0.5526
At 15 docs	0.5263
At 20 docs	0.5026
At 30 docs	0.4632
At 100 docs	0.3179
At 200 docs	0.2400
At 500 docs	0.1610
At 1000 docs	0.0640



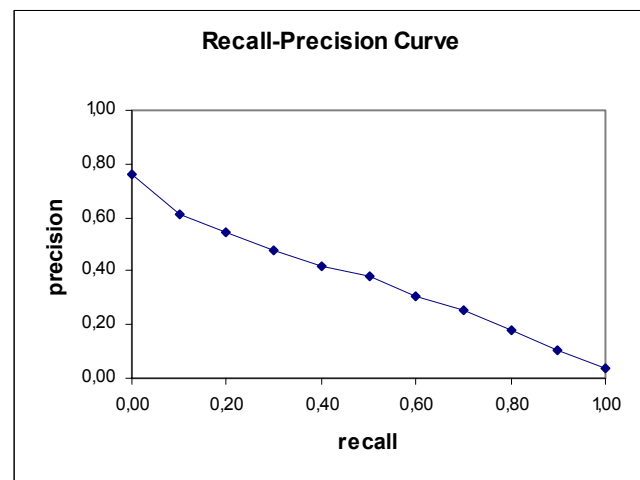
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.4205

(d) Result Set: zh_manual**Summary of Statistics**

Indexing: Chinese Vgram Lexer
 Retrieval: Manual (Boolean and ACCUM operators + user defined search terms)
 Nr. of Topics: 19
 Total Docs retrieved: 5229
 Total relevant Docs: 1399
 Relevant Docs retrieved: 1012

Recall Level Precision Averages

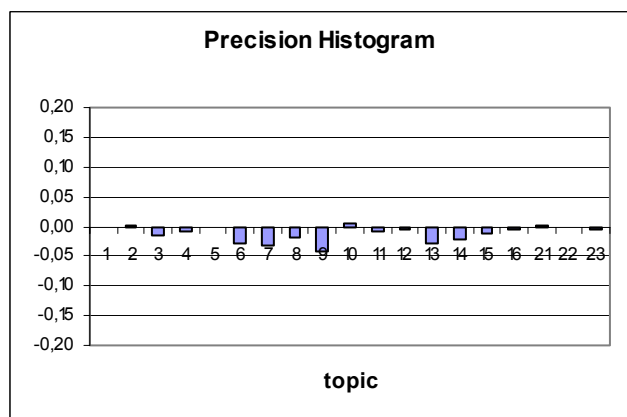
Recall	Precision
0.00	0.7622
0.10	0.6154
0.20	0.5419
0.30	0.4811
0.40	0.4200
0.50	0.3777
0.60	0.3090
0.70	0.2503
0.80	0.1758
0.90	0.1954
1.00	0.0371



Average precision over all relevant docs: 0.4594 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.6526
At 10 docs	0.5895
At 15 docs	0.5333
At 20 docs	0.5079
At 30 docs	0.4702
At 100 docs	0.3137
At 200 docs	0.2497
At 500 docs	0.1353
At 1000 docs	0.0813



R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.4176

D.2 Experiment II

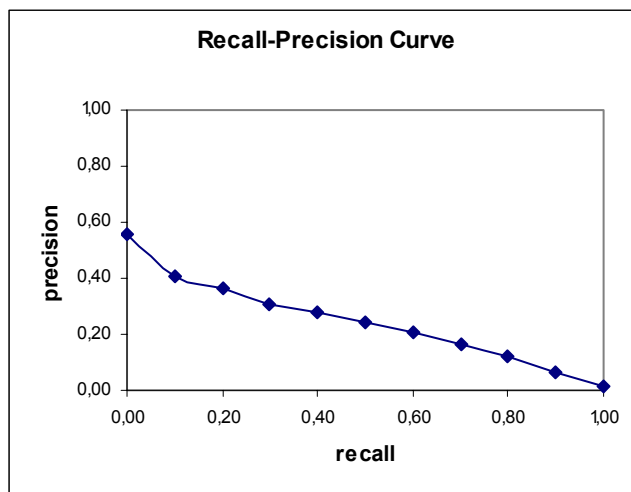
(a) Result Set: ch_auto

Summary of Statistics

Indexing: Chinese Lexer
 Retrieval: Auto (TREC-defined descriptors connected by ACCUM operator)
 Nr. of Topics: 19
 Total Docs retrieved: 17502
 Total relevant Docs: 1399
 Relevant Docs retrieved: 1027

Recall Level Precision Averages

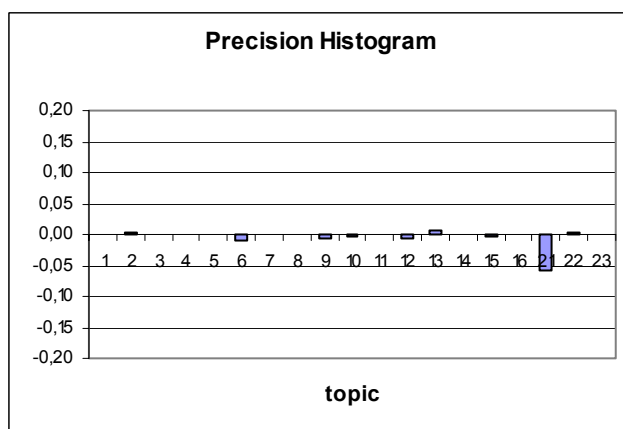
Recall	Precision
0.00	0.5578
0.10	0.4053
0.20	0.3678
0.30	0.3076
0.40	0.2786
0.50	0.2453
0.60	0.2104
0.70	0.1674
0.80	0.1204
0.90	0.0617
1.00	0.0113



Average precision over all relevant docs: 0.2863 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.3789
At 10 docs	0.3842
At 15 docs	0.3649
At 20 docs	0.3579
At 30 docs	0.4772
At 100 docs	0.2311
At 200 docs	0.1634
At 500 docs	0.0927
At 1000 docs	0.0971



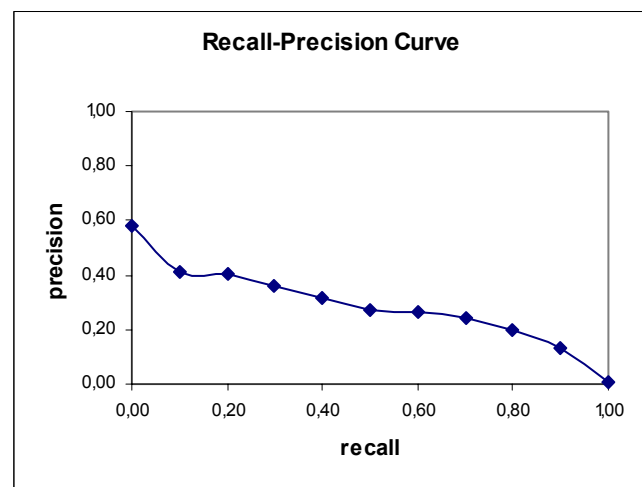
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.2886

(b) Result Set: weighted_auto**Summary of Statistics**

Indexing: Chinese Lexer
 Retrieval: Auto (weighted query terms + ACCUM operator)
 Nr. of Topics: 19
 Total Docs retrieved: 17500
 Total relevant Docs: 1399
 Relevant Docs retrieved: 1175

Recall Level Precision Averages

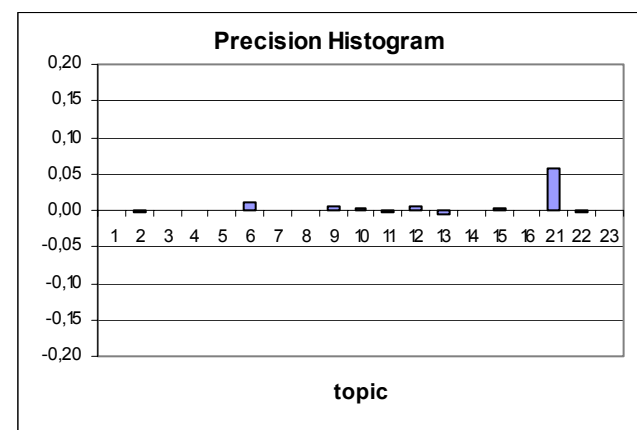
Recall	Precision
0.00	0.5786
0.10	0.4108
0.20	0.4021
0.30	0.3568
0.40	0.3195
0.50	0.2702
0.60	0.2611
0.70	0.2395
0.80	0.1992
0.90	0.1344
1.00	0.0083



Average precision over all relevant docs: 0.3309 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.4316
At 10 docs	0.4000
At 15 docs	0.4175
At 20 docs	0.3974
At 30 docs	0.5500
At 100 docs	0.2684
At 200 docs	0.2042
At 500 docs	0.1072
At 1000 docs	0.0414



R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.3327

D.3 Experiment III

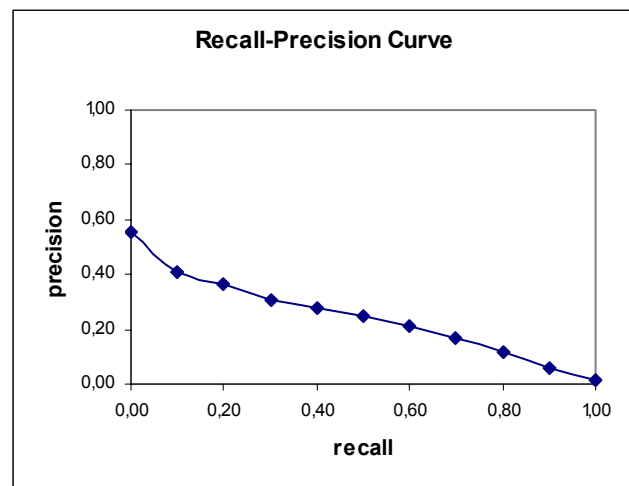
(a) Result Set: stop1_auto, stop2_auto, stop3_auto

Summary of Statistics

Indexing: Chinese Lexer
 Retrieval: Auto (TREC-defined descriptors connected by ACCUM operator)
 Nr. of Topics: 19
 Total Docs retrieved: 17502
 Total relevant Docs: 1399
 Relevant Docs retrieved: 1027

Recall Level Precision Averages

Recall	Precision
0.00	0.5578
0.10	0.4053
0.20	0.3678
0.30	0.3076
0.40	0.2786
0.50	0.2453
0.60	0.2104
0.70	0.1674
0.80	0.1204
0.90	0.0617
1.00	0.0113



Average precision over all relevant docs: 0.2863 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.3789
At 10 docs	0.3842
At 15 docs	0.3649
At 20 docs	0.3579
At 30 docs	0.4772
At 100 docs	0.2311
At 200 docs	0.1634
At 500 docs	0.0927
At 1000 docs	0.0971

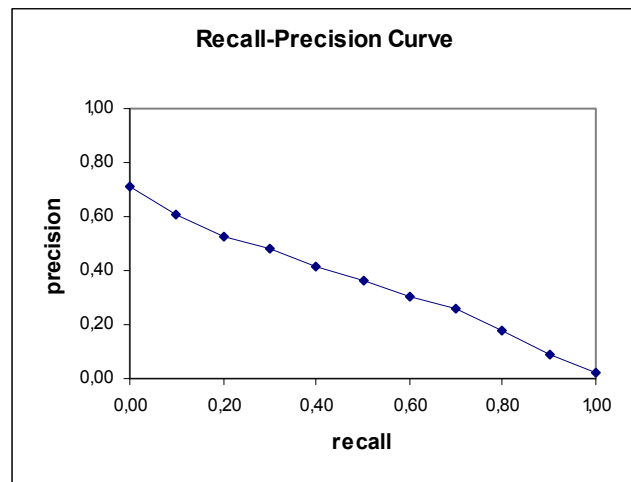
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.2886

(b) Result Set: stop1_manual, stop2_manual, stop3_manual**Summary of Statistics**

Indexing: Chinese Lexer
 Retrieval: Manual (Boolean and ACCUM operators + user defined search terms)
 Nr. of Topics: 19
 Total Docs retrieved: 4280
 Total relevant Docs: 1399
 Relevant Docs retrieved: 951

Recall Level Precision Averages

Recall	Precision
0.00	0.7108
0.10	0.6094
0.20	0.5226
0.30	0.4825
0.40	0.4160
0.50	0.3609
0.60	0.3006
0.70	0.2602
0.80	0.1753
0.90	0.0923
1.00	0.0208



Average precision over all relevant docs: 0.4706 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.6526
At 10 docs	0.5526
At 15 docs	0.5263
At 20 docs	0.5026
At 30 docs	0.4632
At 100 docs	0.3179
At 200 docs	0.2400
At 500 docs	0.1610
At 1000 docs	0.0640

R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.4205

D.4 Experiment IV

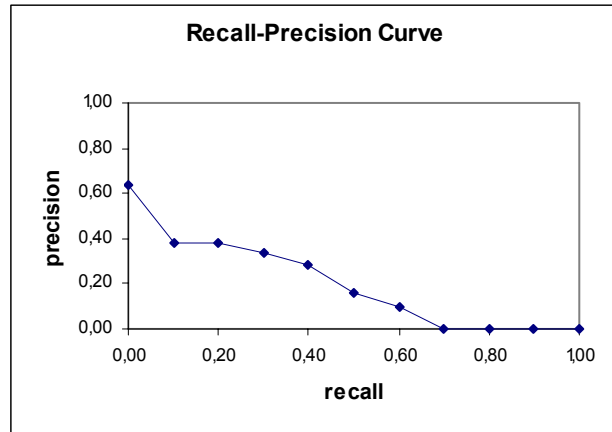
(a) Result Set: zh_thes

Summary of Statistics

Indexing: Chinese Vgram Lexer
 Retrieval: Manual with Thesaurus
 Nr. of Topics: 4
 Total Docs retrieved: 1178
 Total relevant Docs: 212
 Relevant Docs retrieved: 122

Recall Level Precision Averages

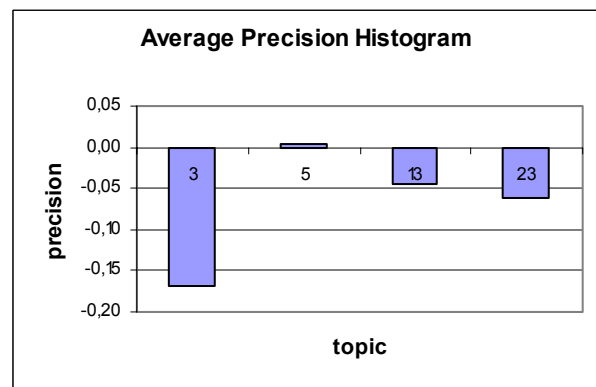
Recall	Precision
0.00	0.6333
0.10	0.3815
0.20	0.3795
0.30	0.3400
0.40	0.2813
0.50	0.1623
0.60	0.0993
0.70	0.0000
0.80	0.0000
0.90	0.0000
1.00	0.0000



Average precision over all relevant docs: 0.3436 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.4500
At 10 docs	0.4500
At 15 docs	0.4167
At 20 docs	0.3625
At 30 docs	0.3250
At 100 docs	0.1950
At 200 docs	0.1425
At 500 docs	0.1300
At 1000 docs	0.0000



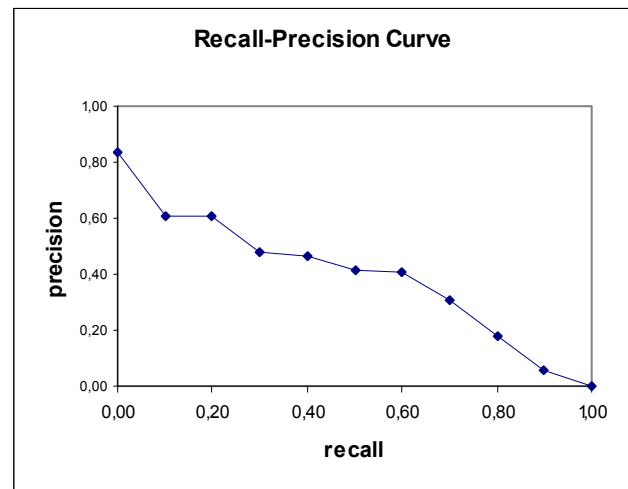
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.2934

(b) Result Set: zh_manual**Summary of Statistics**

Indexing: Chinese Vgram Lexer
 Retrieval: Manual (Boolean and ACCUM operators + user defined search terms)
 Nr. of Topics: 4
 Total Docs retrieved: 1394
 Total relevant Docs: 212
 Relevant Docs retrieved: 179

Recall Level Precision Averages

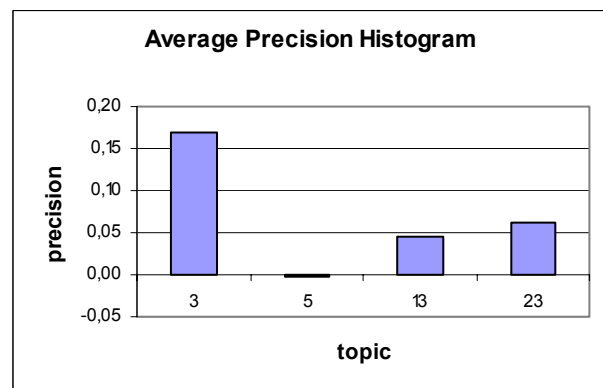
Recall	Precision
0.00	0.8333
0.10	0.6043
0.20	0.6083
0.30	0.4805
0.40	0.4615
0.50	0.4175
0.60	0.4083
0.70	0.3105
0.80	0.1780
0.90	0.0543
1.00	0.0000



Average precision over all relevant docs: 0.4792 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.6526
At 10 docs	0.5526
At 15 docs	0.5263
At 20 docs	0.5026
At 30 docs	0.4632
At 100 docs	0.3179
At 200 docs	0.2400
At 500 docs	0.1610
At 1000 docs	0.0640



R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.4341

D.5 Experiment V

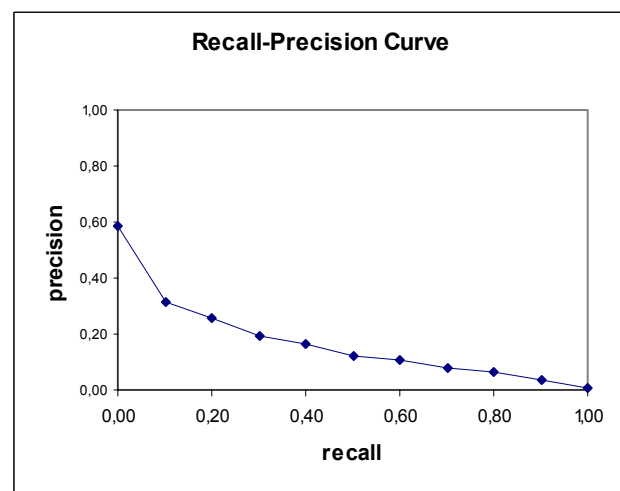
(a) Result Set: User 1

Summary of Statistics

Indexing: Chinese Vgram Lexer
 Retrieval: Manual (user-formed queries)
 Nr. of Topics: 19
 Total Docs retrieved: 11120
 Total relevant Docs: 1399
 Relevant Docs retrieved: 739

Recall Level Precision Averages

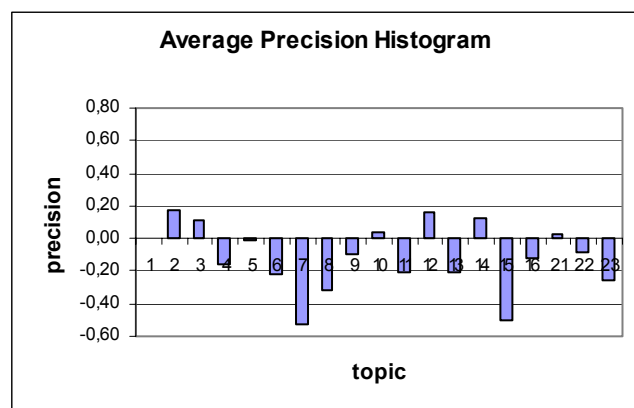
Recall	Precision
0.00	0.5848
0.10	0.3118
0.20	0.2591
0.30	0.1919
0.40	0.1657
0.50	0.1189
0.60	0.1071
0.70	0.0774
0.80	0.0676
0.90	0.0368
1.00	0.0097



Average precision over all relevant docs: 0.3105 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.3778
At 10 docs	0.3000
At 15 docs	0.3059
At 20 docs	0.3313
At 30 docs	0.3021
At 100 docs	0.1947
At 200 docs	0.1504
At 500 docs	0.1008
At 1000 docs	0.0510



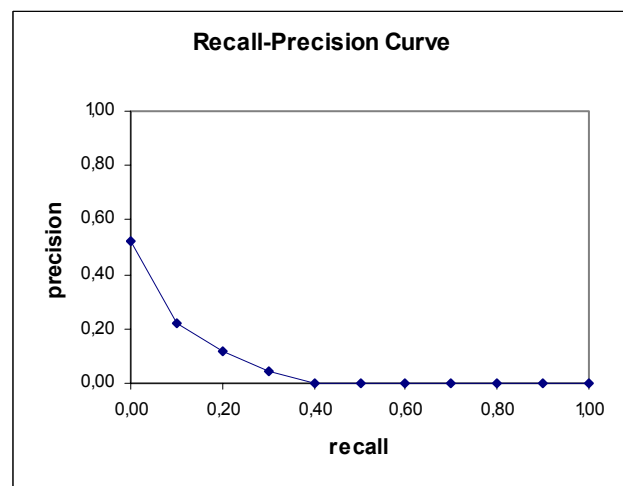
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.2013

(b) Result Set: User 2**Summary of Statistics**

Indexing: Chinese Vgram Lexer
 Retrieval: Manual (user-formed queries)
 Nr. of Topics: 19
 Total Docs retrieved: 503
 Total relevant Docs: 1399
 Relevant Docs retrieved: 152

Recall Level Precision Averages

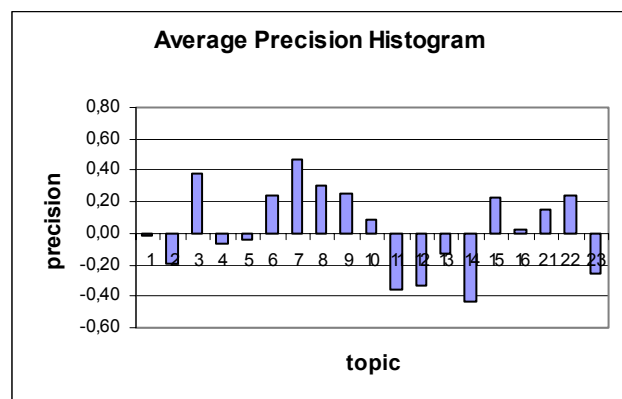
Recall	Precision
0.00	0.5222
0.10	0.2220
0.20	0.1173
0.30	0.0425
0.40	0.0000
0.50	0.0000
0.60	0.0000
0.70	0.0000
0.80	0.0000
0.90	0.0000
1.00	0.0000



Average precision over all relevant docs: 0.4480 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.3714
At 10 docs	0.5143
At 15 docs	0.4952
At 20 docs	0.4833
At 30 docs	0.4222
At 100 docs	0.3050
At 200 docs	0.0000
At 500 docs	0.0000
At 1000 docs	0.0000



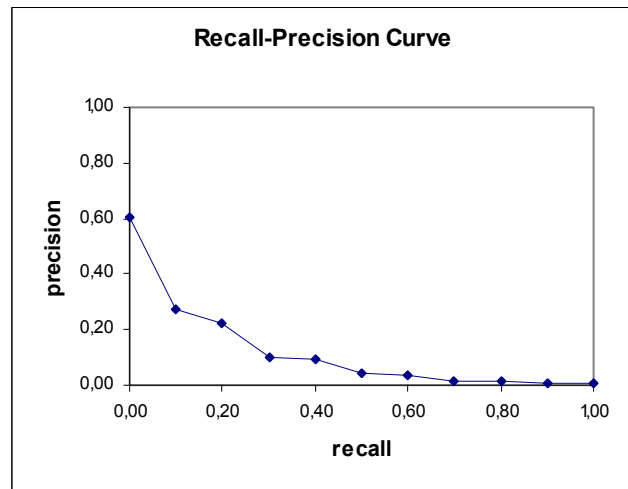
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.0965

(c) Result Set: User 3**Summary of Statistics**

Indexing: Chinese Vgram Lexer
 Retrieval: Manual (user-formed queries)
 Nr. of Topics: 19
 Total Docs retrieved: 3739
 Total relevant Docs: 1399
 Relevant Docs retrieved: 435

Recall Level Precision Averages

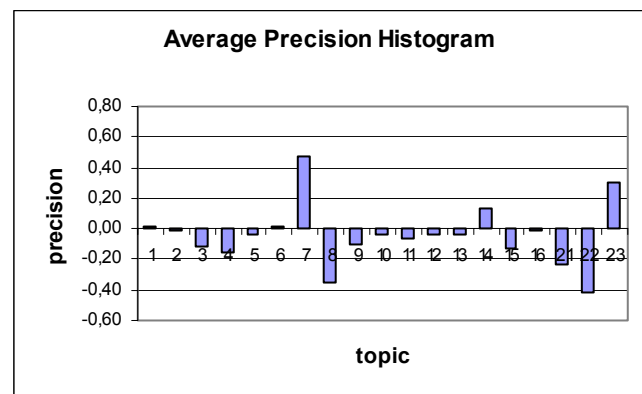
Recall	Precision
0.00	0.6008
0.10	0.2752
0.20	0.2255
0.30	0.1016
0.40	0.0936
0.50	0.0423
0.60	0.0332
0.70	0.0154
0.80	0.0132
0.90	0.0037
1.00	0.0041



Average precision over all relevant docs: 0.3762 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.3750
At 10 docs	0.3538
At 15 docs	0.3846
At 20 docs	0.3792
At 30 docs	0.3750
At 100 docs	0.2191
At 200 docs	0.1736
At 500 docs	0.1180
At 1000 docs	0.0660



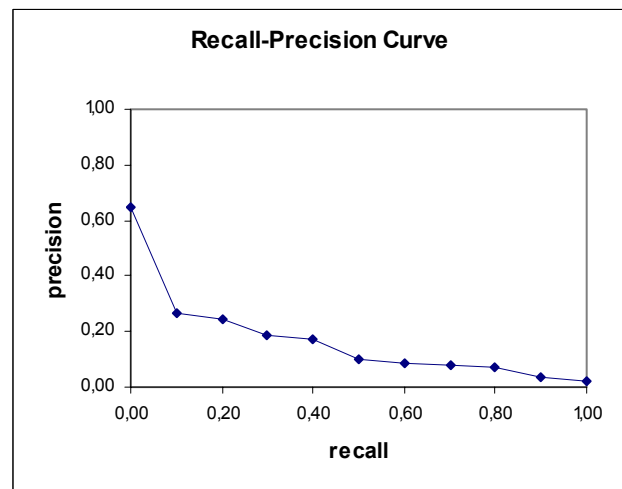
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.1690

(d) Result Set: User 4**Summary of Statistics**

Indexing: Chinese Vgram Lexer
 Retrieval: Manual (user-formed queries)
 Nr. of Topics: 19
 Total Docs retrieved: 8314
 Total relevant Docs: 1399
 Relevant Docs retrieved: 458

Recall Level Precision Averages

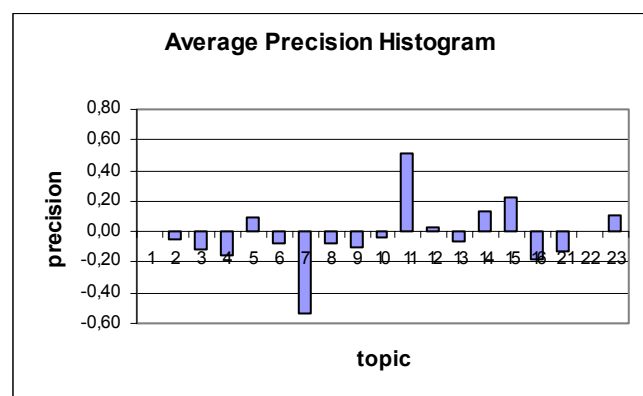
Recall	Precision
0.00	0.6472
0.10	0.2675
0.20	0.2423
0.30	0.1875
0.40	0.1692
0.50	0.1024
0.60	0.0879
0.70	0.0766
0.80	0.0709
0.90	0.0394
1.00	0.0247



Average precision over all relevant docs: 0.3979 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.4000
At 10 docs	0.3857
At 15 docs	0.3667
At 20 docs	0.3393
At 30 docs	0.3026
At 100 docs	0.1915
At 200 docs	0.1461
At 500 docs	0.0636
At 1000 docs	0.0630



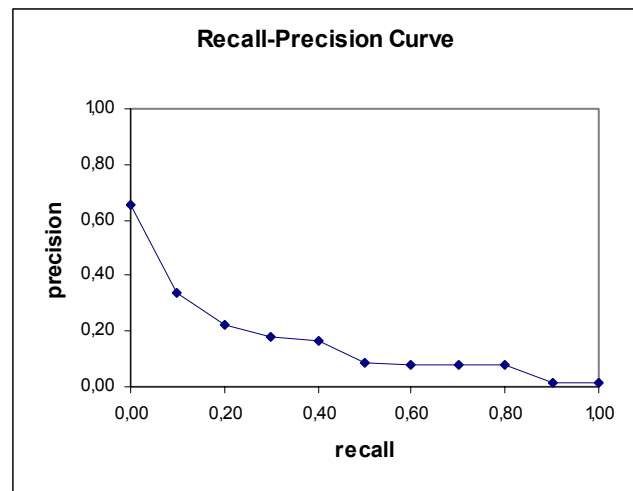
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.1951

(e) Result Set: User 5**Summary of Statistics**

Indexing: Chinese Vgram Lexer
 Retrieval: Manual (user-formed queries)
 Nr. of Topics: 19
 Total Docs retrieved: 1570
 Total relevant Docs: 1399
 Relevant Docs retrieved: 260

Recall Level Precision Averages

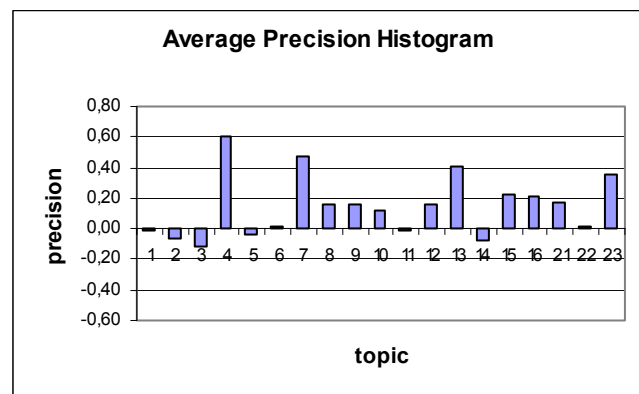
Recall	Precision
0.00	0.6534
0.10	0.3369
0.20	0.2259
0.30	0.1789
0.40	0.1656
0.50	0.0874
0.60	0.0825
0.70	0.0765
0.80	0.0757
0.90	0.0164
1.00	0.0168



Average precision over all relevant docs: 0.5649 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.4444
At 10 docs	0.4643
At 15 docs	0.4476
At 20 docs	0.4600
At 30 docs	0.3889
At 100 docs	0.2413
At 200 docs	0.0650
At 500 docs	0.0370
At 1000 docs	0.0480



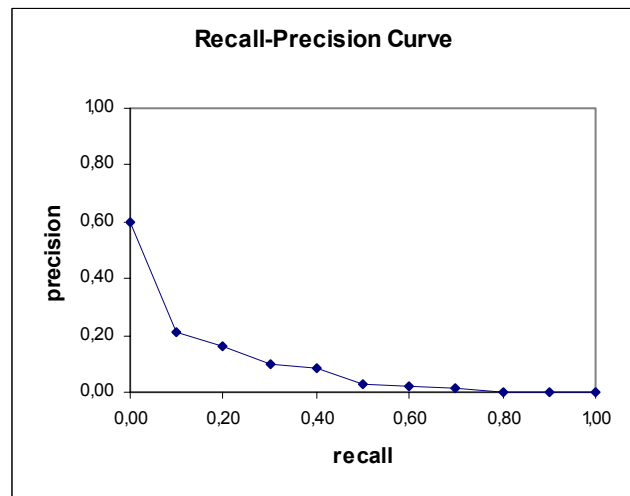
R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.1822

(f) Result Set: User 6**Summary of Statistics**

Indexing: Chinese Vgram Lexer
 Retrieval: Manual (user-formed queries)
 Nr. of Topics: 19
 Total Docs retrieved: 1337
 Total relevant Docs: 1399
 Relevant Docs retrieved: 303

Recall Level Precision Averages

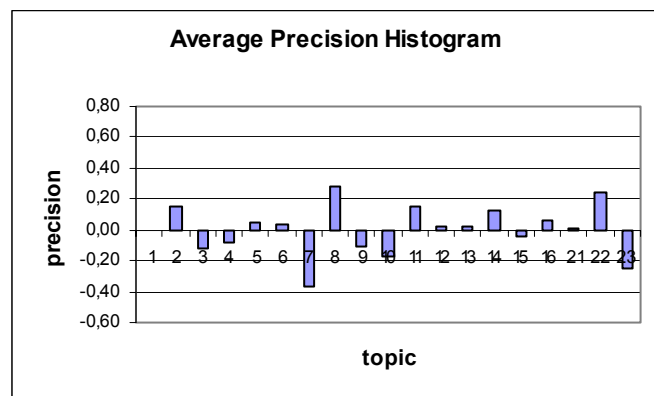
Recall	Precision
0.00	0.5955
0.10	0.2131
0.20	0.1635
0.30	0.0973
0.40	0.0852
0.50	0.0278
0.60	0.0225
0.70	0.0126
0.80	0.0000
0.90	0.0000
1.00	0.0000



Average precision over all relevant docs: 0.4214 (non-interpolated)

Document Level Averages

Level	Precision
At 5 docs	0.3765
At 10 docs	0.4000
At 15 docs	0.4500
At 20 docs	0.4208
At 30 docs	0.3939
At 100 docs	0.2000
At 200 docs	0.2550
At 500 docs	0.1270
At 1000 docs	0.0000



R-Precision (precision after R docs retrieved, where R is the number of relevant documents): 0.1396

Appendix E: Stoplists

E.1 Oracle's Default Stoplist for Simplified Chinese (stop2)

必将	必须	并非	由于	一同	一再	一得
超过	成为	除了	处在	此项	从而	存在着
达到	大量	带来	带着	但是	当时	得到
都是	对于	这个	而且	而言	方面	各方面
各种	共同	还将	还有	很少	很有	还是
回到	获得了	或者	基本上	基于	即可	较大
尽管	就是	具有	可能	可以	来自	两个
之一	没有	目前	哪里	那里	却是	如果
如何	什么	实在	所需	所有	它的	他们
为了	我们	下去	现在	相当	新的	许多
也是	以及	已经	以上	因此	因为	

E.2 Oracle's Default Stoplist for Traditional Chinese

目前	由于	因此	他們	可能	沒有	希望
有關	不過	可以	如果	對於	因為	是否
但是	相當	其中	其它	雖然	我們	包括
必須	以上	之后	所以	以及	許多	最近
至于	一般	不是	不能	而且	引起	如何
除了	不少	最后	就是	分別	加強	甚至
繼續	另外	共同	祇有	了解	根据	已經
過去	所有	不會	以來	任何	一直	不同
立即	左右	經過	尤其	使得	相關	當時
進入	并不	据了解	現在	知識	需要	原因
祇要	否則	并未	什麼	如此	不要	

E.3 Modified Stoplist (stop3)

上	名	以	个	为	和	对	并	版	已	的	等
与	说	完	中	了	在	将	是	电	有	要	

References

- Alpha, S. et al.** (2001): Oracle at TREC-10: Filtering and Question-Answering. In "Proceedings of the Tenth Text Retrieval Conference (TREC-10)", pp. 423-433.
- Baeza-Yates, R.; Ribeiro-Neto, B.** (1999): Modern Information Retrieval. Addison-Wesley, Essex, England.
- Buckland, M.; Gey F.** (1994): The Trade-off between Recall and Precision. *Journal of the American Society for Information Science*, 45(1):12-19.
- Campbell, J.:** Chinese Dialects. URL: <http://www.glossika.com/en/dict/index.htm> (Date of Access: 12.12.2003).
- Chen, J.** (2001): Oracle9i Text Chinese Support (PowerPoint Presentation). Oracle Corp., California.
- Chinese and Oriental Languages Information Processing Society, COLIPS** (中文与东方语文信息处理学会): COLIPS Homepage. URL: <http://www.colips.org> (Date of Access: 10.1.2004).
- Chinese Language Society of Hong Kong** (香港中国语文学会): Huayuqiao (华语桥). URL: <http://huayuqiao.org> (Date of Access: 16.02.2004).
- Fabian, M.** (2003): 中日韩 (CJK) Support in SuSE Linux (electronic copy). URL: <http://www.suse.de/~mfabian/suse-cjk.pdf> (Date of Access: 15.10.2003).
- Feng, Z.W. (冯志伟)** (1988): The Characteristics and Difficulties in Automatic Chinese Semantic Analysis – a Lesson Learned from Chinese-English Machine Translation (从汉英机器翻译看汉语自动句法语义分析的特点和难点). In "Quantitative and Computational Studies of the Chinese Language" (汉语计量与计算研究), Editor T'sou, B.K. et al. LISRC, City University, Hong Kong.
- Harter, S.P.** (1996): Variations in Relevance Assessment and the Measurement of Retrieval Effectiveness. *Journal of the American Society for Information Science*, 47(1):37-49.
- He, J. et al.:** (1996): Berkeley Chinese Information Retrieval at TREC-5: Technical Report. In "Proceedings of the Fifth Text Retrieval Conference (TREC-5)", pp. 191-195.
- Huang, X.; Robertson, S.E.** (1997): Okapi Chinese Text Retrieval Experiments at TREC-6. In "Proceedings of the Sixth Text Retrieval Conference (TREC-6)", pp. 137-142.
- Ji, D.H. (姬东鸿); Huang, C.N. (黄昌宁)** (1996): A Semantic Composition Model for Chinese Nouns and Adjectives (汉语形容词和名词的语义组合模型). *Journal of Chinese Language and Computing*, 6(1):25-33.

- Kekäläinen, J.; Järvelin, K.** (2002): Using Graded Relevance Assessments in IR Evaluation. *Journal of the American Society for Information Science*, 53(13):1120-1129.
- Kwok, K.L.** (1999): Employing Multiple Representations for Chinese Information Retrieval. *Journal of the American Society for Information Science*, 50(8):709-723.
- Kwok, K.L.; Grunfeld, L.** (1996): TREC-5 English and Chinese Retrieval Experiments using PIRCS. In "Proceedings of the Fifth Text Retrieval Conference (TREC-5)", pp. 133-142.
- Kwok, K.L.; Grunfeld, L.; Xu, J.H.** (1997): TREC-6 English and Chinese Retrieval Experiments using PIRCS. In "Proceedings of the Sixth Text Retrieval Conference (TREC-6)", pp. 207-214.
- Leong, M.K.; Zhou, H.** (1997): Preliminary Qualitative Analysis of Segmented vs Bigram Indexing in Chinese. In "Proceedings of the Sixth Text Retrieval Conference (TREC-6)", pp. 551-557.
- Luk, R.W.P.; Kwok, K.L.** (2002): A Comparison of Chinese Document Indexing Strategies and Retrieval Models. *ACM Transactions on Asian Language Information Processing*, 1(3):225-268.
- Lunde, K.** (1999): *CJKV Information Processing*. O'Reilly, California.
- Mahesh, K.; Kud, J.; Dixon, P.** (1999): Oracle at TREC-8: a Lexical Approach. In "Proceedings of the Eighth Text Retrieval Conference (TREC-8)", pp. 207-216.
- Mateev, B. et al.** (1997): ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval. In "Proceedings of the Sixth Text Retrieval Conference (TREC-6)", pp. 623- 636.
- Microsoft® Corporation** (2001): Comparing Windows XP Professional Multilingual Options. In "Microsoft Technet", URL: <http://www.microsoft.com/technet> (Date of Access: 20.02.2004).
- National Institute of Standards and Technology (NIST)**: Text Retrieval Conference Homepage. URL: <http://trec.nist.gov> (Date of Access: 22.01.2004).
- Ngo, C.W.; Lai, K.F.** (1996): Experiments on Routing, Filtering and Chinese Text Retrieval in TREC-5. In "Proceedings of the Fifth Text Retrieval Conference (TREC-5)", pp. 247-255.
- Nie, J.Y.; Brisebois, M.; Ren, X.** (1996): On Chinese Text Retrieval. In "Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval", pp. 225-233.
- Nohr, H.** (2001): *Automatische Indexierung: Einführung in betriebliche Verfahren, Systeme und Anwendungen*. Verl. für Berlin-Brandenburg, Postdam.

- Nohr, H.** (2003): Grundlagen der automatischen Indexierung: ein Lehrbuch. Logos-Verlag, Berlin.
- Oracle Corporation** (2003): *Oracle® Text Reference*, Release 1(10.1), Beta 1. Oracle Corp., California, U.S.A.
- Packard, J.L.** (2000): *The Morphology of Chinese: a Linguistic and Cognitive Approach*. Cambridge University Press, Cambridge (U.K.).
- Rajaraman, K.; Lai, K.F.; Changwen, Y.** (1997): Experiments on Proximity Based Chinese Text Retrieval in TREC 6. In "Proceedings of the Sixth Text Retrieval Conference (TREC-6)", pp. 559-567.
- Salton, G.; McGill, M.** (1983): *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.
- Savage-Knepshield, P.A.; Belkin, N.J.** (1999): Interaction in Information Retrieval: Trends Over Time. *Journal of the American Society for Information Science*, 50(12):1067-1082.
- Schamber, L.** (1994): Relevance and Information Behavior. *Annual Review of Information Science and Technology*, 29:3-48.
- Smeaton, A.; Wilkinson, R.** (1997): Spanish and Chinese Document Retrieval in TREC-5. In "Proceedings of the Fifth Text Retrieval Conference (TREC-5)", pp. 57-64.
- Spink, A.; Saracevic, T.** (1997): Interaction in Information Retrieval: Selection and Effectiveness of Search Terms. *Journal of the American Society for Information Science*, 48(8):741-761.
- Sproat, R.** (2001): *Corpus-Based Methods in Chinese Morphology and Phonology* (Notes for a course presented at the 2001 Summer Institute of the Linguistic Society of America, in the Subinstitute on Chinese Corpus Linguistics at the University of California, Santa Barbara, July 15 – August 3, 2001).
- Su, X.C.** (苏新春) (2000a): Some Findings about Chinese Words from the "Modern Chinese Dictionary" (关于《现代汉语词典》词汇计量研究的思考). In "Chinese Bridge" (华语桥), URL: <http://huayuqiao.org> (Date of Access: 16.02.2004).
- Su, X.C.** (苏新春) (2000b): Homographs and the Semantics of Words – an Analysis of Homographs in the "Modern Chinese Dictionary" (同形词与“词”的意义范围 — 析《现代汉语词典》的同形词词目). In "Chinese Bridge" (华语桥), URL: <http://huayuqiao.org> (Date of Access: 16.02.2004).
- Tague-Sutcliffe, J.M.** (1996): Some Perspectives on the Evaluation of Information Retrieval Systems. *Journal of the American Society for Information Science*, 47(1):1-3.

- Tong, X. et al.** (1996): Experiments on Chinese Text Indexing: CLARIT TREC-5 Chinese Track Report. In "Proceedings of the Fifth Text Retrieval Conference (TREC-5)", pp. 335- 339.
- Van Rijsbergen, C.J.** (1979): Information Retrieval. Department of Computer Science, University of Glasgow.
- Voorhees, Ellen M.** (1998): Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In "Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval", pp. 315-323.
- Wang, H.D. (王惠迪)** (2002a): Chinese Characters, Words and Sentences (华文字词句). Lingzi Media, Singapore.
- Wang, H.D. (王惠迪)** (2002b): Characteristics of Singapore Chinese (新加坡华语特有词语探微). In "Singapore Chinese Words and Grammar" (新加坡华语词汇与语法), Editor Zhou, Q.H. (周清海). Lingzi Media, Singapore.
- Wang, J.** (2003): Chinese Segmentation: a Pragmatic View. In "Chinese Syntax and Semantics", Editors. Xu, J.; Ji, D.; Lua, K.T. Prentice Hall, Singapore.
- Wilkinson, R.** (1998): Chinese Document Retrieval at TREC-6. In "Proceedings of the Sixth Text Retrieval Conference (TREC-6)", pp. 25-29.
- Yale University, Council of East Asian Studies:** Chinese Mac Frequently Asked Question (Website). URL: <http://www.yale.edu/chinesemac/index.html> (Date of Access: 20.02.2004).
- Zhou, Q.H. (周清海); Xiao, G.Z. (萧国政)** (1998): Forms, Semantics and Use of Chinese Words in Singapore (新加坡华语词的形选择、词义选择和词用选择). In "Proceedings of the International Conference on Teaching of Mandarin" (普通话 (国语) 教学国际研讨会), Macau, 11-13 March 1998.

Acknowledgments

I would like to extend my grateful thanks to my supervisors Professor Dr.-Ing. Peter Lehmann and Frau Barbara Steinhanses for their guidance and enthusiasm. Without their initiation and joint-efforts, this project would not have been possible.

A sincere thank you also to the colleagues of Oracle Deutschland GmbH who have rendered considerable amount of help in the initial stages of database set up. They are, in alphabetical order, Stephan Bohn, Jörg Eggelsmann, Tom Robert, Jürgen Vester and Harald Wiedemer.

I am grateful to Paul Dixon and Ciya Liao of the *Oracle® Text* development team for giving invaluable feedback and comments on my work, and Ole Olsen of *Digital Collections*, as well as Reiner Ebenhöch and May Chua of *Atex Media Command (ACM)* for sharing the Chinese thesaurus and newspaper corpus of the *Singapore Press Holdings*.

Besides, I sincerely thank Professor Kim-Teng Lua of the *National University of Singapore* for inviting me to help out at the *17th Pacific Conference on Language, Information and Computation (PACLIC)*. This event has proven to be very useful for acquiring a better understanding of the issues related to text segmentation and information retrieval in the Chinese language.

In addition, I want to thank my friends Daniel Aw-Yong, Hui Ling Chy, Eunice Lee, Szer Ming Lee, Min Huey Ong and Kong Weng Wong for their participation in Experiment V of the thesis.

Finally, special thanks to my family and friends for their concern, and to my husband Ralf Scharnweber for his encouragement and continuous support, and for proof-reading this thesis for me.