「情報管理研究」, vol. 23, no. 1, 1992. 3., pp. 39～62

國內 文獻情報 檢索을 위한 키워드 自動抽出 시스템 開發

Automatic Keyword Extraction System for Korean Documents Information Retrieval

芮 龍 熙*
(Yae, Yong Hee)

抄 錄

本 硏究는 실제의 데이터 分析을 통하여 60여개의 助詞와 출현빈도는 높지만 檢索에 불필요한 320여개의 不用語를 선정하여 좌우절단을 적용한 네 가지 유형으로 분류하고 助詞와 불용어 테이블을 구성하는 方法을 제시한다. 한글문헌에서 單語가 추출되면 조사의 효율적인 절단이 이루어지고, 漢字語인 경우 한글로 변환되며, 2단계로 不用語除去 과정을 거쳐 키워드를 선정하는 시스템을 개발한다. 여기서 추출된 키워드는 情報專門家에 의해 추출된 索引語와는 92.2%의 일치율을 보였다. 그리고 4～6글자로 구성된 複合語의 경우 本 硏究에서 제시한 분리방법에 의해 약 2배의 새로운 單語를 추가할 수 있었으며 그 중 58.8%가 키워드로 적합했다.

키 워 드

키워드 自動抽出, 情報檢索, 한글문헌정보, 조사절단, 불용어 제거, 漢字變換, 複合語 分離, 自動索引.

ABSTRACT

In this paper about 60 auxiliary words and 320 stopwords are selected from analysis of sample data, four types of stopword are classified left, right and auxiliary word truncation & normal. And a keyword extraction system is suggested which undertakes efficient truncation of auxiliary word from words, conversion of Chinese word to Korean and exclusion of stopword. The selected keyeords in this system show 92.2% of accordance ratio compared

* 産業技術情報院 電算室 責任硏究員.
Computer System Division, Korea Institute of Industry & Technology Information

with manually selected keywords by expert. And then compound words consist of 4~6 character generate twice of additional new words and 58.8% words of those are useful as keyword.

KEYWORDS

Automatic Keyword extraction, Information retrieval, Korean document, Auxiliary word truncation, Chinese character translation, Compound word, Automatic indexing.

Ⅰ. 序　　論

情報化時代를 맞이하여 國內의 각종 정보 발생량이 폭발적이라고 표현할만큼 급격한 추세를 보이고 있다. 컴퓨터와 데이터통신의 급속한 발전과 보급으로 이런 방대한 양의 정보를 적절히 가공, 축적하고 컴퓨터에 체계적으로 수록하여 정보요구자가 필요로 하는 情報를 신속, 정확하게 찾아내는 일련의 활용을 情報檢索이라 한다.

최근 정보의 양적 증가와 컴퓨터의 처리 능력의 발달로 인한 대규모 데이터베이스가 기계가독형으로 축적되고 있다. 이러한 컴퓨터에 축적된 정보를 대상으로 종래의 수작업에 의한 索引語 선정 대신 컴퓨터에 의해서 索引語를 추출하는 自動索引 技法이 개발되고 있다. 自動索引은 單語別로 띄어쓰기 형식을 취하고 있는 歐美語를 중심으로 많이 연구되었으나, 韓國語의 문장 구조상 그대로 적용하기는 매우 어려운 실정이다. 本 硏究에서는 情報檢索을 위한 국내 文獻情報에서 한글의 특성을 적용한 키워드 自動抽出 과정을 연구하여 情報檢索 效率을 극대화하고자 한다.

이를 위하여 약 8萬件의 國內文獻에서 나타나는 조사들의 빈도를 조사하고, 각 單語의 유형을 분석하여, 情報檢索시스템에 사용될 키워드를 自動抽出하는 과정을 정립한다. 여기에 사용된 자료로는 産業技術情報院에 소장하고 있는 單行本 및 報告書 3萬 7,920件의 題目, 1988년도 定期刊行物 記事索引에 수록된 4萬 2,002件의 제목 및 1988년도 博士學位 論文抄錄 1,368件의 제목과 초록 등이다.

Ⅱ. 情報檢索 시스템

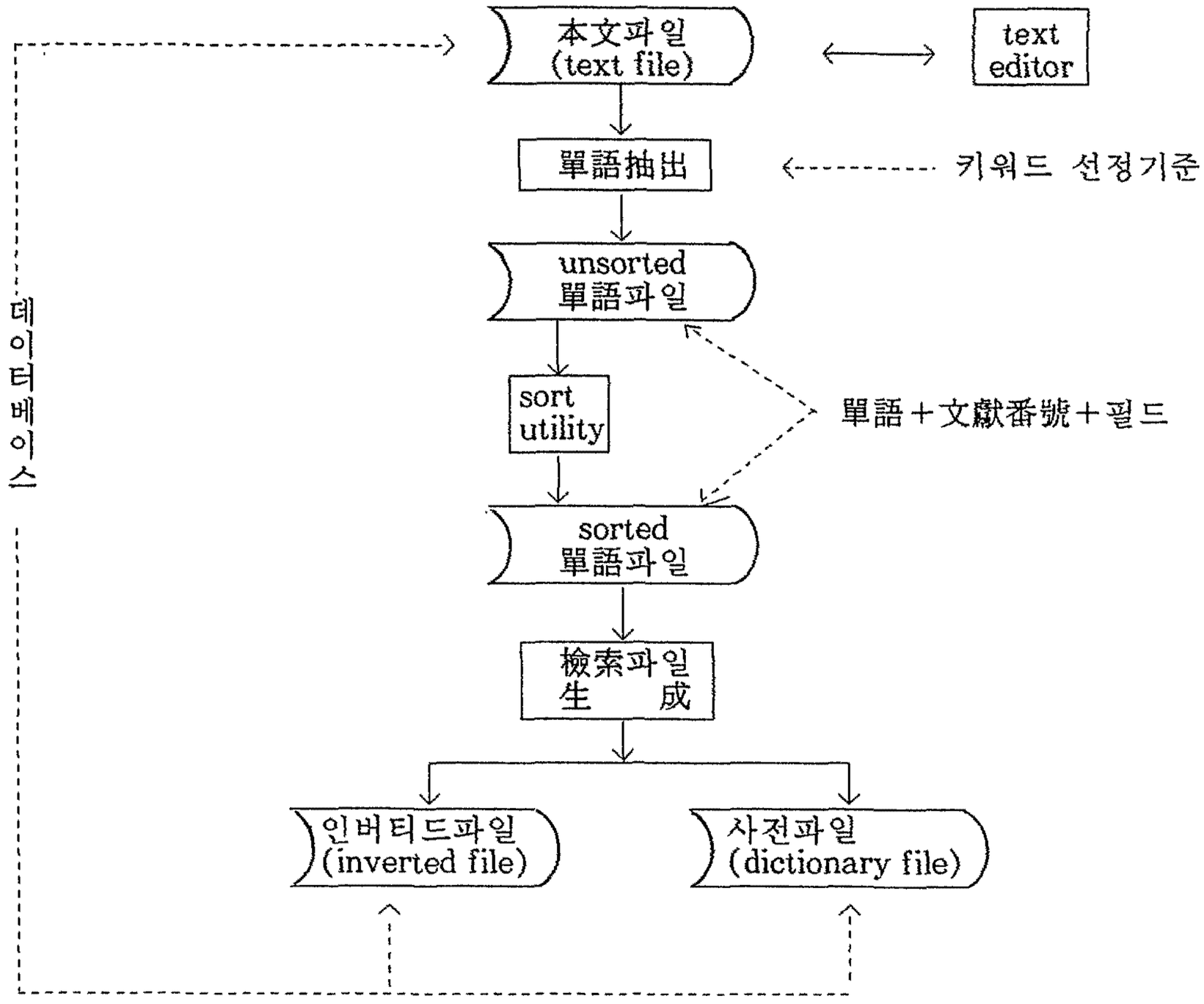
1. 情報의 蓄積

情報檢索 시스템은 入力資料를 파일로 보관하는 本文生成시스템, 檢索을 위한 索引生成 시스템, 필요로 하는 요구사항을 만족하기 위한 情報檢索機能으로 나눌 수 있다.

(1) 本文生成 시스템

本文生成 시스템은 보통 워드프로세서가 가지고 있는 本文編輯(text editor) 機能을 수행할 수 있도록 삽입, 수정, 삭제, 복사이동 등이 가능해야 하며 화면의 내용이 전후방향으로 이동될 수 있어야 한다. 각 文獻을 구성하여 檢索 및 出力의 기본단위로 동작하는 여러개의 패러그래프(혹은 필드)로 나누어져

〈圖 1〉 文獻情報 데이터베이스 構築過程



야 한다. 文獻의 길이와 각각의 패러그래프 길이가 모두 가변길이이므로 대용량의 情報가 저장되는 기억공간을 최소화하도록 해야 한다.

(2) 索引生成 시스템

本文이 만들어지고 나면 檢索되어야 하는 패러그래프가 선정된다. 첫 단계로 그 속에 포함된 單語를 추출하여 文獻番號와 패러그래프 번호를 붙여 적절히 가공, 분석한 후 單語파일을 만든다. 둘째 단계에서 만들어진 단어파일은 sort 유틸리티를 이용하여 알파벳 혹은 가나다순으로 배열한다. 셋째 단계에서 재배열된 단어파일을 읽어 동일한 단어는 한 단어만 남겨 사전파일을 구성하고, 본문을 가리키는 포인터(문헌번호)만을 개수만큼 묶어서 인버티드 파일(inverted file)을 만드는 〈圖 1〉과 같은 索引生成過程을 거치게 된다. 첫 단계의 單語를 추출하는 과정에서 한글에 관한 특성을 고려하여 여러가지 규칙을 적용시킨다.

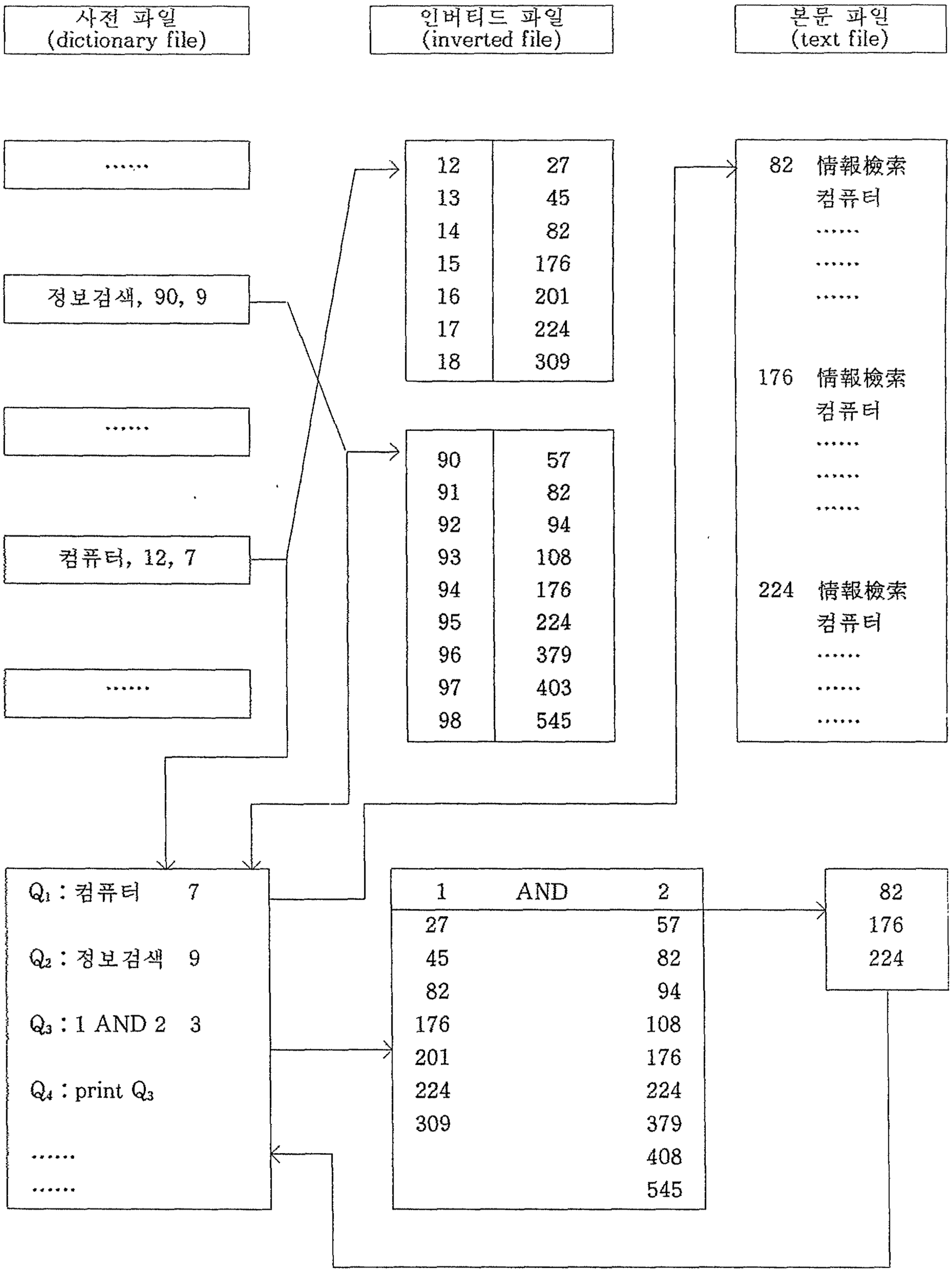
2. 情報檢索原理

情報檢索은 탐색기능, 제한기능, 출력기능 등 몇 가지의 주요한 기능을 가간다. 탐색기능은 AND, OR, NOT 등의 논리연산자와 우측절단 기능을 이용하여 검색하고자 하는 주제를 넓히거나 좁혀간다. 제한기능은 주로 탐색된 결과를 EQ, LT, GT 등 비교연산자를 이용하여 순서, 크기 및 분야의 범위를 제한할 수 있다. 이와 같은 과정을 이용자의 요구사항이 만족될 때까지 질문을 계속하여 최종결과를 얻었을 경우 본문을 화면에 비추거나 프린트에 출력한다.

情報檢索 시스템의 파일구성은 키워드를 사용하여 탐색기능을 갖는 사전파일, 논리연산자를 이용하여 AND, OR, NOT 등의 기능을 수행할 수 있는 인버터 파일, 결과에 대한 문헌을 출력하는데 사용될 본문파일이 있다. 이들 파일들의 구성관계는 〈圖 2〉와 같다.

〈圖 2〉에서 먼저 탐색자가 명령 1에서 "컴퓨터"라는 키워드를 입력하게 되면, 사전파일에서 컴퓨터라는 單語를 읽어 요구조건에 맞는 7건의 문헌이 있다는 것을 알려준다. 명령 2에서 "정보검색"이라는 키워드도 동일한 방법으로 9건이 있다는 것을 알려준다. 명령 3에서 "1 AND 2"로 명령 1과 명령 2에 대한 논리연산자 AND 조합을 요구하게 되면, 각 명령의 결과에 해당하는 인버티드 파일을 읽어 서로 文獻番號를 비교하여 동일한 文獻番號를 가진 3건(82,

〈圖 2〉 情報檢索 파일의 構成 및 檢索原理



176, 224)의 결과를 나타낸다. 그 결과 문헌번호를 키(key)로 하여 본문파일을 읽어 그에 해당하는 내용을 하나씩 화면 혹은 프린트에 출력한다.

3. 情報檢索 시스템 評價

情報檢索 시스템의 성능을 평가할 때 기준이 되는 것은 무엇보다도 그 시스템에 의해서 출력되는 檢索結果가 정보 요구자에게 얼마만큼 만족을 주느냐에 달려 있다. 즉, 정보요구에 대한 이용자의 만족도를 측정하여 情報檢索시스템의 성능을 평가하게 되는 것이다.

이용자의 만족도 측정은 일반적으로 情報檢索 시스템의 檢索效率, 신속성, 경제성의 세 가지 측면에서 측정될 수 있다. 그밖에 응답시간, 이용자의 노력, 출력형식, 관련문헌의 소장범위 등이 포함된다.

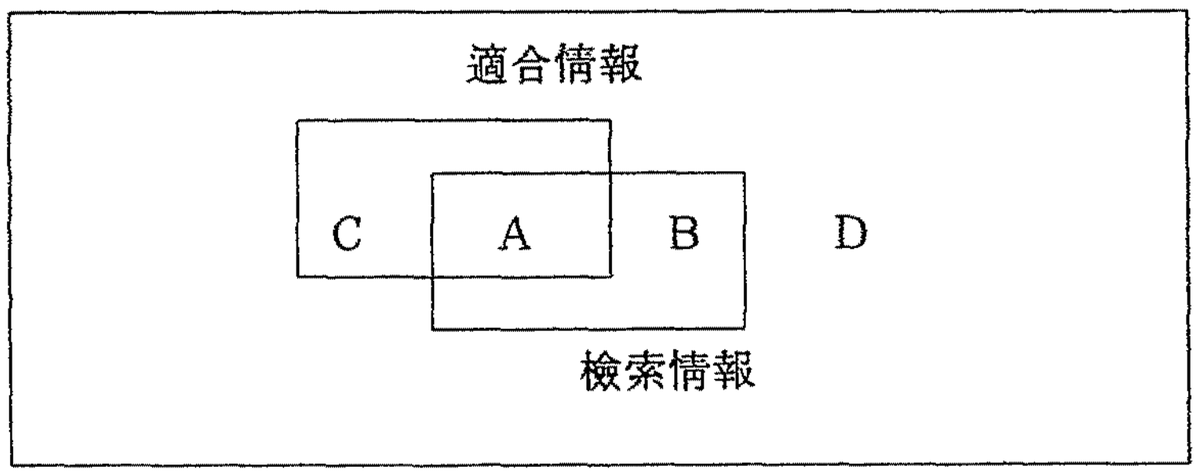
情報檢索을 수행하는 이용자의 궁극적인 목적이 이용자가 필요로 하는 적합문헌을 검색하는데 있기 때문에 여러 평가기준 가운데 檢索效率의 측정방법이 가장 중요한 평가기준이 되고 있다. 檢索效率은 이용자의 정보요구에 적합한 문헌을 검색하는 檢索 시스템의 능력을 의미하는 것으로 검색된 적합 문헌과 부적합문헌, 검색되지 않은 적합문헌과 부적합 문헌 사이의 비율로서 측정된다.

시스템이 소장하고 있는 데이터베이스 내의 본문 파일을 대상으로 情報檢索을 수행하면 전체 본문 파일은 〈圖 3〉과 같이 4개의 집단으로 구분된다. 檢索效率의 측정척도는 再現率(recall ratio)과 精度率(precision ratio)이 널리 사용되고 있다.

〈圖 3〉 情報檢索의 재현율 및 정도율 관계

$$재현율 = \frac{檢索된\ 適合文獻數(A)}{適合文獻\ 總數(A+C)} \times 100$$

$$정도율 = \frac{檢索된\ 適合文獻數(A)}{檢索된\ 文獻總數(A+B)} \times 100$$



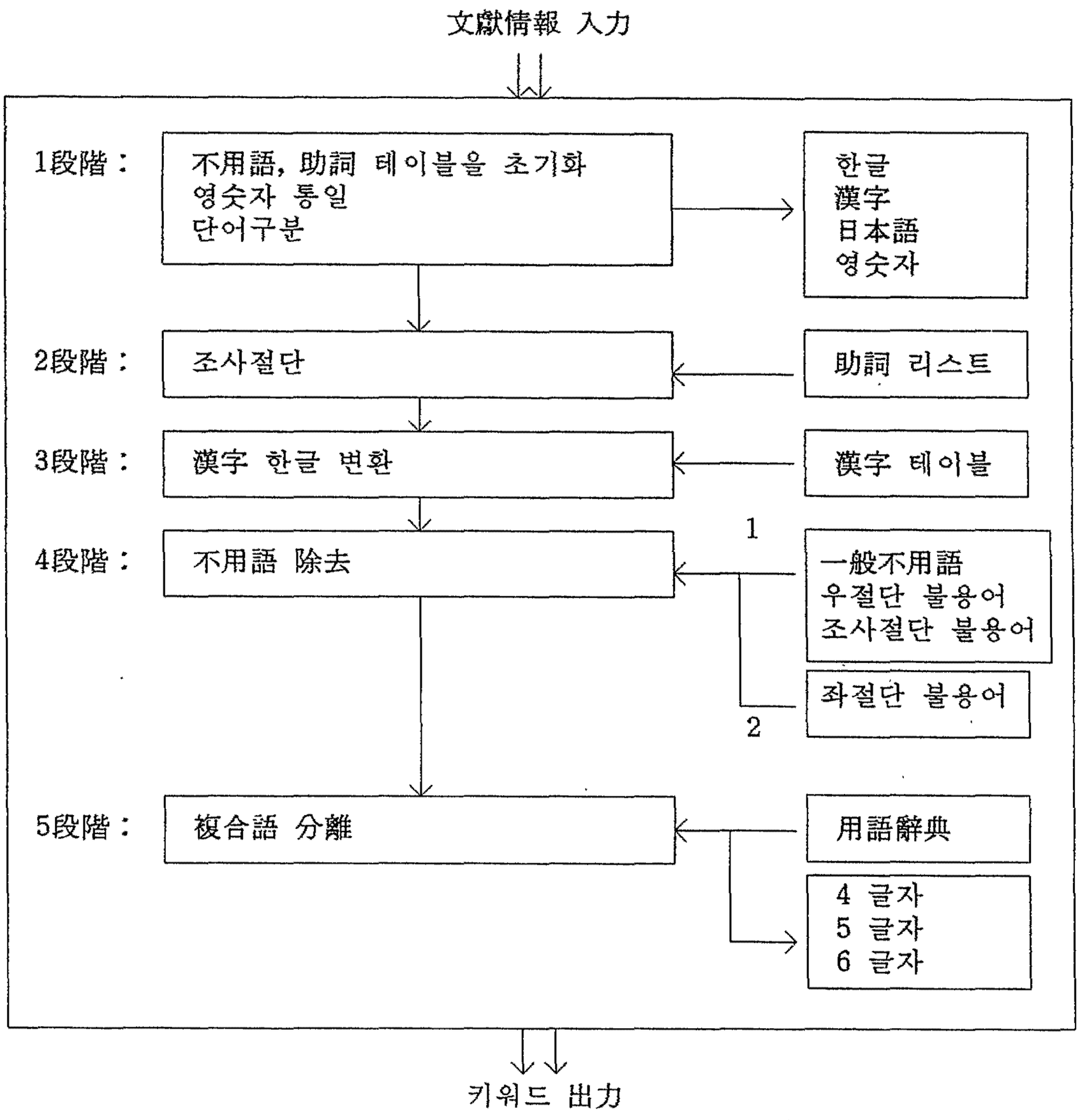
Ⅲ. 키워드 自動抽出 시스템

1. 키워드 抽出 節次

不用語와 助詞 리스트를 읽어 테이블을 만들고, 테이블을 효율적으로 비교하기 위해 포인터 테이블을 작성한다.

1단계로 本文에서 文獻을 읽어 檢索에 필요한 필드를 선정하여 필드의 내용 중 2byte로 구성된 영숫자 및 특수부호는 1byte로 바꾸어 통일하고, 日本語와 더불어 사용된 장음표시 "ー"記號는 그대로 둔다. 필드 내에서 單語의

〈圖 4〉 키워드 自動抽出 시스템 構成



구분은 영숫자로된 單語와 한글 한자로 된 단어를 먼저 구분하고, 공백을 중심으로 단어를 추출한다. 단어는 좌우에 특수문자가 있으면 제거한다. 單語의 종류는 한글, 漢字, 日本語, 영숫자로 구분한다.

영숫자로 구성된 단어일 경우 종래의 방법인 전치사, 접속사, 대명사 등으로 이루어진 100여개의 불용어를 제외한 모든 單語를 키워드로 선정하고, "-"가 연결된 복합단어는 추가로 "-"를 중심으로 단어를 분리시킨다. 나머지 종류의 단어는 2단계에서 5단계의 과정을 반복하면서 키워드를 추출한다. 키워드 自動抽出 과정을 요약하면 〈圖 4〉와 같다.

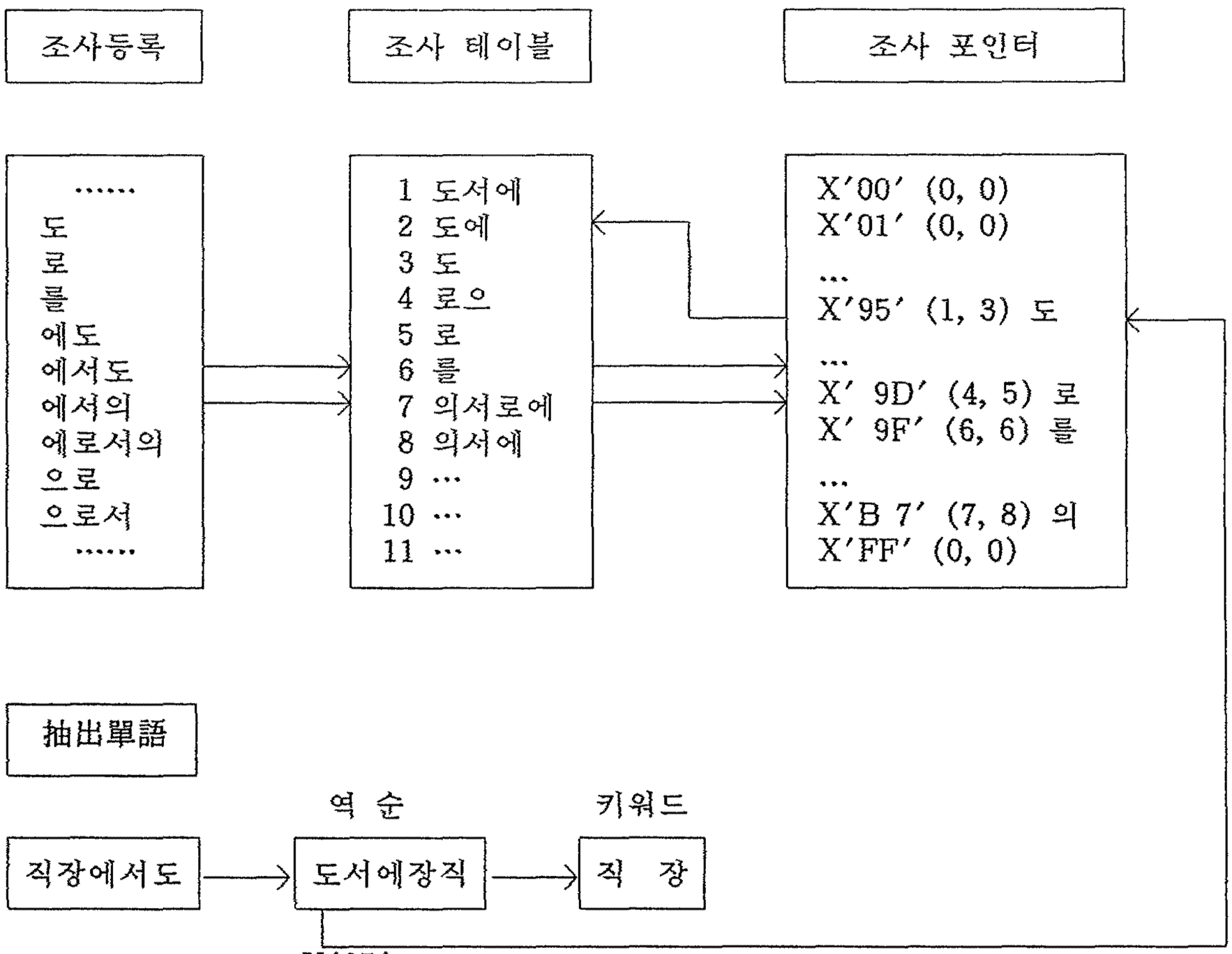
2. 助 詞 切 斷

본문에서 추출된 單語의 끝 부분에 조사가 사용되었는지를 알기 위해서는 單語의 뒤에서부터 앞으로 한글자씩 진행하면서 조사 테이블의 내용과 비교하여야 한다. 이때 스택을 이용한 최장일치의 원칙에 의해 조사를 찾는다. 예를 들면, "직장으로부터"라는 어절에서 "으로부터", "로부터", "부터" 모두가 조사 테이블에 있을 경우 음절의 길이가 가장 긴 "으로부터"가 조사로 인식되어 절단되고 "직장"이 키워드로 남는다.

추출된 단어의 길이가 M자이고, 조사 테이블의 크기가 N일 경우 순차적으로 비교하면 比較回數가 M×N번이나 되어 비효율적이다. 本 硏究에서는 조사 테이블에 있는 조사들을 글자의 역순으로 배열하여 첫 글자의 가나다순과 조사의 길이가 긴순으로 배열시켜 조사 테이블을 만들고 첫 글자의 2byte 중 첫번째 1byte의 코드 값이 같은 조사 테이블의 하한값(low bound)과 상한값(upper bound)을 갖는 0-255의 조사 테이블을 구성한다.

본문에서 單語가 抽出되면 글자의 역순으로 배열하여 첫글자의 첫 byte의 코드값을 구하여 조사 포인터 테이블에 登錄된 조사 테이블의 범위를 확인하여 값이 (0,0)이면 單語의 끝부분이 조사가 아니라는 것을 의미하고, 값이 (a, b)이면 조사테이블의 a에서 b 사이를 순차비교하여 조사를 검출한다. 〈圖 5〉에서 "직장에서도"라는 단어는 "도서에장직"의 역순으로 배열하여 "도"의 첫 1byte의 값이 X'95'이므로 조사 포인터 테이블에서 값 (1, 3)을 얻는다. 조사 테이블의 1에서 3번째 사이를 순차비교하여 "도서에"가 조사로 검출되어 절단되고, "직장"만이 키워드로 남게 된다. 조사의 개수가 60여개이고, 256개의 조사 포인터 테이블 중 18개만이 조사 테이블을 가리키고 있어 평균 조사 3개 정도가 포인터 테이블에 할당되어 평균 비교횟수는 거의 1회이다.

〈圖 5〉 조사 테이블 構成 및 檢出過程

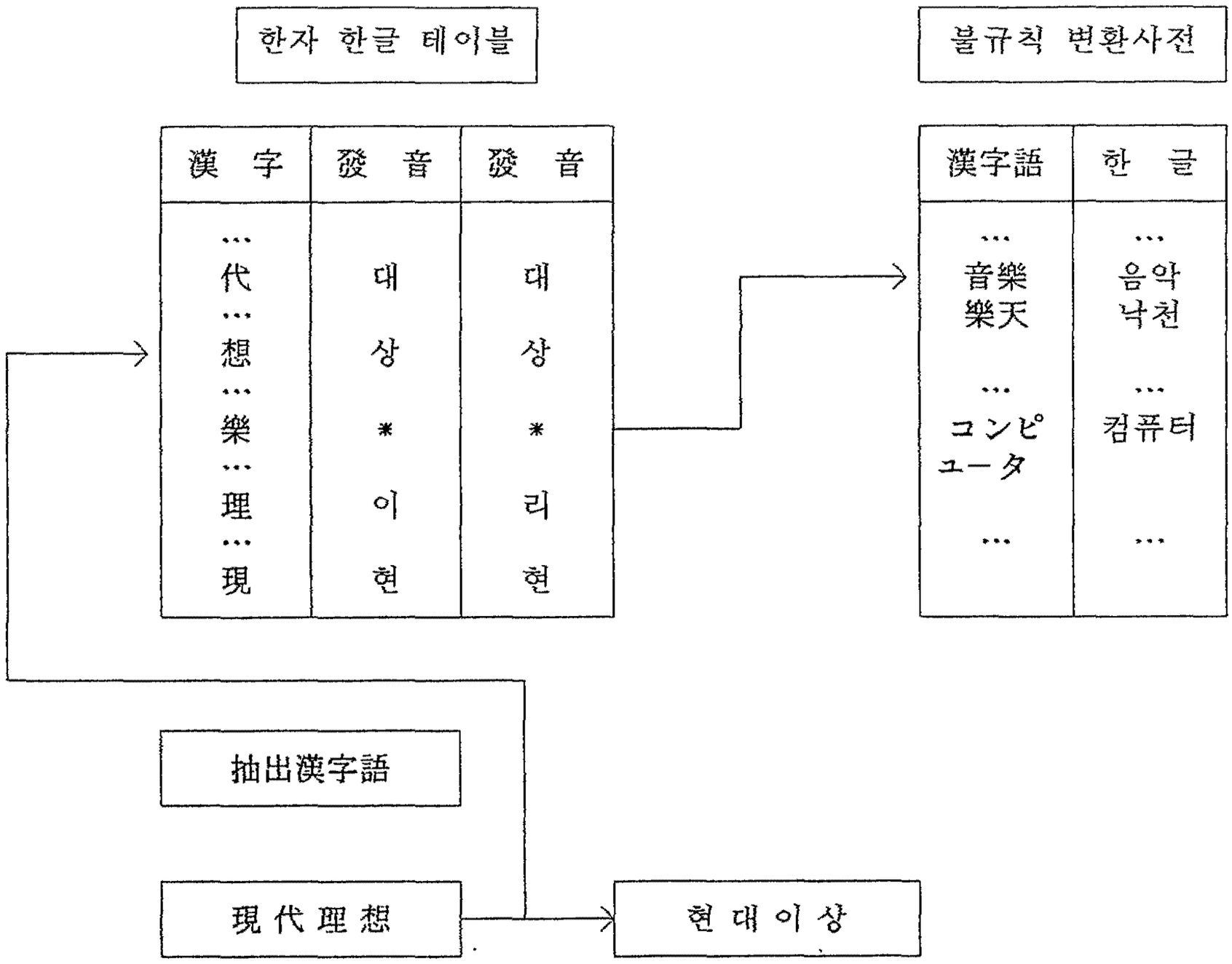


3. 한자 한글 변환

검색용 探索言語는 키 보드로 쉽게 入力이 가능한 한글과 영숫자가 되는 것을 원칙으로 한다. 産業技術情報院의 所藏資料에서 영어단어를 제외한 키워드 중 單行本은 74.3%가, 기사색인은 43.8%, 學位論文抄錄은 42.4%가 漢字語를 사용하고 있다. 본문에서 추출된 단어가 漢字語이면 한글로 변환해야 하며, 그러기 위해서는 한글변환 사전이 필요하게 된다. 모든 한자 단어를 수록한 사전을 만들기는 어려우며 컴퓨터에 사용되는 漢字 약 5,000자의 글자를 대상으로 한글 테이블을 작성한다.

한자 한글 변환은 문법적인 요소들이 많이 적용된다. 문법의 특성인 두음법칙, 동자이음 등의 문제들이 사용하는 言語體系에 상반됨이 없이 처리하여야 한다. 예를 들면, "車"자에 있어서 "自動車"와 "自轉車"는 "자동차"와 "자전거"로 발음되고, "樂"자는 "악", "낙", "락", "요" 등으로 불규칙하게 발음되

〈圖 6〉 한자한글 변환 시스템



되고, "理"자의 경우는 "理想"은 "이상"으로, "物理"는 "물리"로 두음법칙에 따라 발음된다.

그러므로 한자에 대한 한글 테이블을 작성할 때 〈圖 6〉과 같이 두음법칙에 의한 발음과 정상발음의 2가지를 만들어 추출된 한자어의 마지막 글자가 홀수인 경우를 제외하고, 단어의 홀수번째는 두음법칙의 발음을 적용하고 짝수번째는 정상발음을 적용시킨다. 두음법칙을 제외한 불규칙 변환일 경우는 테이블에 "*"를 표시해 두어 日本語 등과 더불어 불규칙 변환에 대한 한글변환 사전을 만들어 예외적인 처리를 한다.

4. 不用語 제거

不用語(stopword)란 情報檢索에서 探索言語로 사용되지 않으면서 문헌 내 출현빈도가 높은 단어들을 말한다. 일반적으로 英語에서는 모든 單語가 띄어쓰기 형식을 취하기 때문에 代名詞, 前置詞, 接續詞 등을 불용어로 취급하고

있다. 그러나 한글은 띄어쓰기는 하고 있지만 단어에 다양한 종류의 조사 및 어미가 붙어 있고, 複合語와 動詞의 활용이 다양하여 불용어의 선택에 어려움이 많다.

리스버겐에 의하면 불용어를 사용하여 중요하지 않은 단어들을 제거함으로써 檢索시스템 전체 문헌파일의 크기를 30~50%까지 줄일 수 있다. 그러나 불용어를 과다하게 선정할 경우에는 索引語로 선정되어야 할 단어가 제외되고 너무 적게 선정할 경우에는 불필요한 단어가 索引語로 선정되므로 불용어 선정에 신중을 기해야 한다.

本 硏究에서는 韓國科學技術情報센터에서 정한 100여개의 불용어와 國內文獻에 나타나 있는 불용어를 중심으로 실험 데이터와 비교하여 추가 또는 삭제하였다. 또한 實驗 데이터에서 나타난 單語의 글자를 역순으로 배열하여 용언의 어미가 활용하는 유형 및 빈도를 조사하여 "-하는", "-되는" 등과 같은 좌측절단(left truncation) 불용어 55개를 선정하였다. 또한 어미 변화에 따른 불용어수의 증가를 효율적으로 처리하기 위해 우측절단(right truncation) 방법을 많이 사용하고 있다. 즉, "관계"라는 단어가 불용어로 될 경우 "관계*"으로 표시해 주면 "관계를", "관계있는" 등의 단어들이 불용어로 간주되는 것을 말하는데, 한글의 명사나 대명사일 경우 우측절단 기법을 사용하면 "관계"라는 말에 다른 名詞가 연결되어 複合名詞인 "관계대명사" 등도 불용어로 된다. 이에 본 硏究에서는 우측절단 기법을 변형한 조사절단 불용어를 적용시켜 "*"를 "%"로 대치시켜 "관계%"로 불용어를 등록함으로써 "관계" 다음에 조사가 이어질 경우 우측절단 기법을 적용하고, 名詞가 이어져 複合名詞로 되는 단어가 키워드가 되도록 한다(〈부록1〉 참조).

불용어의 구분은 다음 네 가지 종류로 분류한다.

첫째, "위하여", "낮은" 등과 같은 일반적으로 자주 사용되는 단어(")

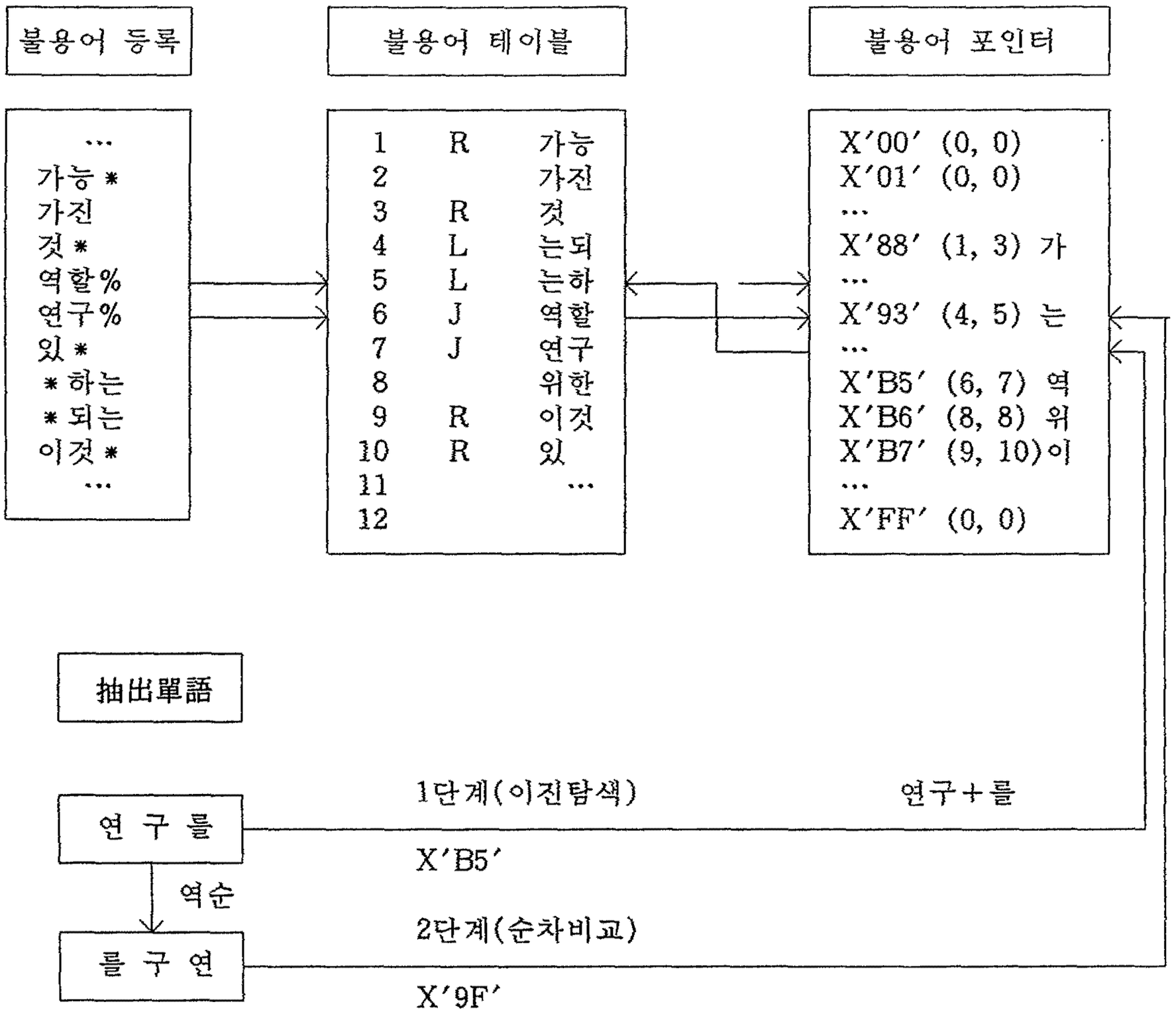
둘째, "않-", "것-" 등과 같이 우측에 어떠한 단어가 이어져도 키워드가 될 수 없는 우절단 불용어('R')

셋째, "목적-", "연구-" 등과 같은 名詞가 조사테이블에 있는 조사만으로 이어지는 조사절단 불용어('J')

넷째, "-있는", "-하는" 등과 같은 용언의 어미가 활용하는 좌절단 불용어('L')

여기서 좌절단 불용어는 조사와 마찬가지로 글자를 역순으로 나열시키고,

〈圖 7〉 불용어 테이블 構成 및 處理節次



전체 불용어의 가나다순으로 불용어 테이블을 재구성한 후 첫 글자의 2byte 중 첫번째 1byte의 코드 값이 같은 불용어 테이블의 하한값과 상한값을 갖는 0~255의 불용어 포인터 테이블을 구성한다.

單語가 불용어인가를 확인하기 위해서는 〈圖 7〉과 같이 2단계를 거쳐 처리되는데, 1단계로 조사를 절단한 후 첫 글자의 1byte의 코드 값을 구하여 불용어 테이블의 범위를 확인한다. 값이 (0, 0)이면 불용어가 아니라는 것을 의미하고, 값이 (a, b)이면 불용어 테이블의 a와 b 사이를 이진탐색(binary search) 방법으로 찾는다. 발견되면 불용어인데, 불용어 구분이 조사절단 불용어인 경우 단어에 조사가 연결되었으면 불용어로 처리하고, 불용어가 아니면 2단계로 조사를 절단하지 않은 상태에서 단어의 글자를 역순으로 배열한 후 첫글자의 첫 1byte의 값을 조사비교 방법과 거의 같은 방법으로 좌절단 불용어 중에서

순차적으로 비교하게 된다. 〈圖 7〉의 예에서 "연구를"에서 조사를 절단한 후 "연구"의 "연"자를 첫 1byte의 코드 값이 X'B5'로 불용어 포인터 테이블에서 값(6, 7)을 구한다. 불용어 테이블의 6과 7번째를 이진탐색방식으로 "J연구"를 찾게 된다. 여기서 "를"이라는 조사가 연결되고, 불용어 유형이 "J"를 나타내므로 조사절단 불용어로 인식되어 불용어로 제거된다. 만약 불용어로 처리되지 않으면 다음 단계로 조사와 더불어 "를구연"의 역순으로 배열하여 좌절단 불용어 중에서 한번 더 비교를 하게 된다.

5. 複合語 分離

한글은 단어와 단어가 연결되어 하나의 단어가 되고, 띄어쓰기의 오류 등으로 인한 複合單語가 많다. 이런 複合單語가 하나의 단어로 취급되어서는 정확한 情報檢索이 이루어지지 않는다. 예를 들면, "데이터通信", "通信裝備" 등은 "通信"이라는 단어로는 동시에 檢索이 되지 않고 "＊通信＊"과 같이 좌우절단 방법을 이용하여야만 檢索이 가능하다. 이런 방법보다는 "데이터", "通信", "데이터通信"과 "通信", "裝備", "通信裝備"의 키워드로 분리하여 索引語 파일을 구성하고, 이용자는 이중 어느 한 單語로 檢索을 할 수도 있고, 檢索質問式에서 複合語를 효율적으로 검색해 내는 방법을 고려할 수도 있다.

이러한 복합어 분리는 좌우측에서 한글자씩 절단하면서 나머지 단어를 용어사전파일에 존재여부를 확인하여 키워드로 선정할 수도 있다. 본 연구에서는 용어사전을 갖고 있지 않으므로 複合語를 분리하기 전의 상태에서 만들어진 3萬여 단어를 갖는 키워드 파일을 용어사전 파일로 사용한다. 키워드 중 6글자 이하가 전체 단어의 96.4%이므로 복합어의 기준을 4~6글자로 구성된 단어로 가정하여 분리할 때 모든 경우의 수를 적용하는 것보다 비현실적인 분리를 제외한 〈表 1〉과 같은 방법으로 분리시킨다.

〈表 1〉 複合語 글자수에 따른 분리 글자수

複合語글자수	분 리 글 자 수	備 考
4 자	1/3, 3/1, 2/2	- 부분의 분리 단어만 용어 사전에서 확인
5 자	2/3, 3/2	
6 자	4/2, 2/4, 3/3, 2/2/2	

Ⅳ. 結 果 分 析

1. 一致率 計算

情報檢索을 위한 키워드 자동추출 시스템의 평가는 情報檢索 시스템의 검색 결과에 의해 평가될 수 있으나 많은 양의 정보를 대상으로 적합한 문헌이 검색되었는지 평가하기는 어렵다.

검색에 불필요한 불용어를 제외하고 키워드를 자동추출하기 때문에 비교적 많은 단어가 키워드로 선정되었으나, 專門家에 의한 수작업으로 추출된 키워드가 자동추출 키워드에 포함되어 있으면 정보검색시 주제에 따라 檢索語를 적절히 조합함으로써 원하는 문헌에 접근할 수 있다.

그러므로 본 논문에서는 Salton이 제시한 유사계수 공식을 변형한 일치율 공식을 평가의 척도로 이용했으며, 복합어를 분리하기 전단계에서 계산한다. 그 공식은 다음과 같다.

$$\text{일치율}(\%)=\frac{B}{A}\times 100$$

A : 主題 專門家에 의해 수작업에서 추출된 색인어수

B : 자동추출된 키워드 중 수작업 색인어와 일치하는 수

〈表 2〉는 실험대상문헌 29건을 대상으로 情報專門家의 수작업에 의한 색인어 추출 도움을 받아 컴퓨터에 의한 자동추출 키워드와 비교분석한다. 문헌 각각에 대한 일치율을 계산하였으며, 전체 문헌에 대한 평균 일치율은 92.2%로 매우 높게 나타나고 있다

2. 漢字語 한글변환

조사 포인터 테이블을 이용하여 비교적 빠른 속도로 조사절단이 이루어졌으며, 제목보다 초록에서 조사의 출현빈도가 높았고 그 중 "새마을", "一효과" 등과 같은 조사어로 끝나는 단어에 조사가 생략된 경우에 오류가 발생되었다. 한자어가 많이 포함되어 있어 큰 문제는 되지 않았으나, 한글자료인 경우 이를 해결하기 위해서는 용어사전 파일에 조사어로 끝나는 단어들을 등록하여

〈表 2〉 수작업과 자동 키워드 추출과의 일치율

文獻番號	총 출 현 키 워 드 수	총 출 현 불 용 어 수	unique 키 워 드 수	수 작 업 키워드수A	일 치 수 B	일 치 율 (%)
1	65	44	44	6	6	100
2	78	70	62	16	14	87.5
3	68	69	44	11	9	81.8
4	65	74	46	13	12	92.3
5	40	39	52	13	12	92.3
6	72	85	46	10	9	90
7	42	89	28	7	7	100
8	73	75	47	16	16	100
9	62	66	42	10	10	100
10	91	97	71	5	3	60
11	53	63	36	7	7	100
12	62	78	41	7	7	100
13	78	68	56	5	4	80
14	83	52	66	9	8	88.9
15	50	77	39	12	12	100
16	74	58	54	11	9	81.8
17	63	68	44	12	12	100
18	75	94	63	6	6	100
19	66	61	57	7	7	100
20	57	53	54	7	7	100
21	58	62	43	8	7	87.5
22	60	66	34	5	5	100
23	38	54	27	5	4	80
24	65	88	37	7	7	100
25	48	78	32	7	6	85.7
26	51	75	39	7	6	85.7
27	64	64	43	5	2	40
28	73	59	61	11	11	100
29	70	62	50	10	10	100
合計	1,874	1,988	1,358	255	235	92.2
(%)		(51.5)				

조사를 절단하기 전 확인과정을 거치므로 해결할 수 있다.

4. 불용어 제거

한글 情報檢索 시스템에 맞는 완벽한 불용어 리스트를 당장 만드는 것은 거의 불가능한 일이며, 기본적인 불용어를 대상으로 자료의 분석과 검색과정을 반복하면서 분야와 내용에 맞는 불용어 리스트를 계속 보완해 나가는 것이 바람직하다.

어미 활용이 발달한 韓國語의 특성을 고려하여 자료분석 결과 "-하는", "-되는" 등의 유형으로 끝나는 좌측절단 불용어를 선정함으로써 많은 양의 불용어를 등록한 효과를 나타냈으며, 내용규명에 관계가 없는 명사에 조사절단불용어를 적용하므로 불용어와 키워드의 구분이 〈表 3〉과 같이 명확하게 분리되었다.

특히, 내용규명에 관계가 없는 단어 34개를 선정하여 전체 대상자료에서 조사절단 불용어를 적용한 결과 〈表 4〉와 같이 18.2%가 名詞로 연결된 複合語로서 키워드로 선정되었다. 만약, 조사절단 불용어를 등록하지 않을 경우 〈表 5〉에서와 같이 전체 불용어의 15.4%가 키워드로 선정되어 색인파일이 커지고 적합정보를 찾는 檢索效率이 나빠지게 된다.

〈表 3〉 조사절단에 의한 불용어와 키워드

키 워 드	불 용 어	키 워 드	불 용 어
관계개념에	관계를	개발계획을	개발
관계개념을	관계	개발권양도제	개발에
관계구문	관계가	개발규모	개발하기
관계대명사와	관계에	개발도는	개발과
관계모형	관계가	개발도상국의	개발이
관계법령	관계로	개발모형	개발하거나
관계변수간의	관계에서	개발문제를	개발된
관계변화	.....	개발방향의	개발을
관계사		개발수입	
관계설정의		개발수입효과	
관계성		개발이익이	
.....		개발정책	

〈表 4〉 조사절단 불용어의 출현비율

	單 行 本	記事索引	學位論文	合 計
조사절단 불용어 수	7,903	11,476	2,951	22,230 (81.8%)
키 워 드 수	452	3,298	1,212	4,962 (18.2%)
合 計	8,385	14,774	4,163	27,292

〈表 5〉 불용어 類型別 出現頻度

	일반불용어 (' ')	우절단 불용어 ('R')	좌절단 불용어 ('L')	조사절단 불용어 ('J')	合 計
출현회수	358	625	699	306	1,988
比 率(%)	18.0	31.4	35.2	15.4	100

本 실험결과 불용어의 比率은 51.7%로 나타났으며, 그 중 조사 유형별로 보면 〈表 5〉와 같이 조사 및 좌절단 불용어가 전체의 50.6%를 차지하고 있다. 이로 말마암아 용언의 어미활용을 통해 많은 양의 불용어를 登錄한 효과를 거둘 수 있다.

5. 複合語 分離

漢字語를 많이 사용하거나, 띄어쓰기가 불분명하여 대체로 複合語가 많이 사용되었으며, 그 複合語를 분리한 결과 〈表 6〉에서와 같이 약 2배 정도의 새로운 單語가 생겨 적어도 한번은 분리되었고, 분리된 단어 중 58.8%가 키워드로 적합하였다. 이로 인해 情報檢索 시스템의 檢索效率을 높일 수 있으나 다만 8% 정도가 전혀 내용과 다른 의미의 단어로 분리되어 이에 대해서는 複合語에 대한 탐색연산자를 檢索 시스템에 적용함으로써 해결할 수 있다.

V. 結 論

情報檢索 시스템에서 索引을 통한 탐색언어는 정보요구자가 필요로 하는 情報에 신속정확하게 접근할 수 있으며, 원하는 情報의 존재여부를 확인해 주는

중요한 역할을 담당한다.

情報檢索 시스템은 국내외적으로 연구가 많이 되어 이용되고 있다. 그러나 키워드 抽出方法이 대부분 英語에서 사용하는 방법을 벗어나지 못하여 情報의 蓄積量이 많아지고, 이용빈도가 높아질수록 情報檢索 효율이 떨어진다는 것을 실감하게 될 것이다. 조사 및 어미를 윗말에 붙여 쓰고 용언의 활용이 발달한 한글문장 構造와 特性이 英語와는 다르고, 단어의 연결로 인한 複合語를 많이 사용하고 있다는 점이 키워드를 추출하는데 어려움을 주고 있다.

근래에 와서 한글문헌의 自動索引에 관한 연구가 점차 중요성이 높아가고 있으나 컴퓨터 한글처리와 사용코드의 차이 등으로 어려움이 많으며, 外來語 및 漢字語 등 표기의 표준화가 이루어지지 않고 있으며, 複合語의 분리문제는 제기되지도 못하고 있는 실정이다.

본 연구에서는 실제로 구축된 많은 양의 자료를 대상으로 데이터를 분석하여, 실험단계에 있는 自動索引에 대한 연구는 효율적인 조사절단, 漢字의 한글변환, 좌우절단을 통한 불용어제거, 복합어분리 등은 어느정도 情報檢索 시스템에서 실용화 될 수 있는 계기가 되고 한글용어사전(thesaurus)이 만들어지고 탐색연산자가 한글특성에 맞게 연구된다면, 情報센터, 硏究所 및 大學 등의 文獻情報處理에 좋은 밑거름이 될 것이다.

〈參 考 文 獻〉

- 科學技術處, 「硏究人力資料 데이터베이스에 관한 硏究」, 서울 : 韓國科學技術情報센터, 1981.
- 곽종우, 「高校國文法」, 서울 : 東亞出版社, 1982.
- 김영환, "한글 한자 혼용문의 자동색인 시스템", 韓國科學技術院 碩士學位論文, 1982.
- 남기심, 고영근, 「標準國語文法論」, 서울 : 탑출판사, 1989.
- 사공철, 「情報檢索論」, 서울 : 아세아출판사, 1985.
- 사공철 외, 「最新情報檢索論」, 서울 : 구미무역(주)출판사, 1990.
- 안현수, "한글문헌의 자동색인에 관한 실험적 연구", 「情報管理學會誌」, vo. 3, no. 2, 1986, pp. 108~306.
- 우동진, "統計的 技法에 의한 한글 자동색인 연구", 「情報管理學會誌」, vol 4, no.1, 1987, pp. 47~86.
- Lancaster, F. W, 「情報檢索시스템」, 윤구호, 김태승 역, 서울 : 구미무역(주)출판사,

1985.

- 이영주, "자동색인을 위한 韓國語 형태소 분석 알고리즘", 「1989년도 한글 및 한국어 정보처리 학술발표 논문집」, 1989, pp. 240~246.
- 정영미, 「情報檢索論」, 서울 : 정음사, 1986.
- 최윤희, 「우리말 情報處理 시스템에 관한 硏究」, 延世大學校 碩士學位論文, 1982.
- 홍종화, 「초성테이블을 이용한 情報檢索시스템 설계 및 구현」, 漢陽大學校 碩士學位論文, 1986.
- 高野文雄外, "日本語キーワード自動抽出ジステム(JAKAS)", 「第18回情報科學技術硏究集會發表論文集」, 1982, pp. 35~44.
- 池原悟 外, "キーワード自動抽出ジステム(INDEXER)", 「硏究實用化報告」, vol. 36, no. 9, 1987, pp. 1,151~1,158.
- Kimoto, H., et al, "Automatic Indexing System for Japanese Text", *Review of the Electrical Communications Laboratories*, vol. 37, no. 1, 1989, pp. 51~56.
- Van Rijsbergen, C. J., *Information Retrival*, 2nd ed., Londan : Butterworths, 1979.
- Jonak, Z., "Automatic Indexing of Full Texts", *Information Processing and Management*, vol. 20, no. 5 ; no. 6, 1984, pp. 619~627.

<附錄 1> 불용어 리스트

가는
가능*
가설*
가운데*
가장
가져*
가진
가질*
각각*
각국*
각기
각종
간섭*
강조*
갖*
같*
개념%
개발%
개별%
개선%
거의
거쳐
걸쳐
검토%
것*
견해%
결과*
결국:
결론*
결정%
경우%
경향*
경험%
계속*
계획%
고찰*
곧
과정*
관계%
관련*
관심%
관점%
관하여
구별*
구체:
규명*
그것*
그대로
그동안
그들*
그래서
그러나
그러면서
그러므로
그런
그런데
그럼에도
그렇지
그에
그의
그이
그중*
근거*
근본%
기능%
기본%
기초%
깊은
까닭*
까지*
끝*
끼친*
나누어*
나아가*
나타나*
나타난
나타내*
남긴
낮은
내용%
내지
넘어*
년
논문%
높*
놓*
다루*
다른
다시
다음*
단순히
달리
대부분
대상%
대하여*
대해*
더불어
더욱이
더욱
동시에
동안*
되*
두가지
둘째*
들어*
따라*
따른
때*
또
또는
많*
맞추어
매우
먼저*
몇*
모두
모든
목적%
목차
목표%
무엇%
문제%
물론*
미치는*
미친
및
바로
바탕*
받*
발전%
밝*
방법%
방안%
방향%
배경%
범위
벗어*
변화*
보는
보다*
보려*
보아
보았*
보여*
보이%
본
본논문%
본연구%
본질%
볼
분석%
비료%
비로소
비록
비추어
뿐*
사실%
사용%
살펴*
새로운
서로
서론
성립%
셋째
속%
수
쉽*
스스로
실제%
실증%
아니라
아닌
아무런
아무리
아울러
아직도
않*
알맞*
어느*
어디까지*
어떤
어렵*
얻
없*
여기%
여러*
역할%
연구%
영향%
오늘날
왔*
왜
왜냐하면
요약%
우리%
우선
위치*
위하여
위해*
의미%

의의 *
의하여
의해 *
이것 *
이는
이때
이런
이론%
이루 *
이르 *
이를
이에
이와 *
이용하 *
이후 *
일
있 *
잘
장
장%
적용%
전개 *
절
점%
접근 *
정도%
제
주는
주로
주요
주체%
중심 *
중요 *
즉
지금 *
지닌
지찾 *
첫째 *
체계 *
추진면 *
측
층
커다란
크기%
큰
타당성%
특성 *
특징 *
특히
파악 *
평가%
포함 *
표시 *
필요 *
하나 *
하는 *
하도록
하며 *
하면서
하여 *
하였 *
한가지
한다 *
한편
할때
할수
함%
함께
합리 *
항
했 *
효과%
후
* 가는
* 게
* 고
* 고자
* 까지
* 높은
* 는데
* 다
* 다는
* 다면
* 도록
* 되거나
* 되기
* 되는
* 되면
* 되어
* 되지
* 된
* 든지
* 때
* 라도
* 려는
* 로써
* 며
* 면서
* 므로
* 수록
* 시켜
* 시키는
* 어야
* 었는
* 었던
* 였던
* 으면
* 있는
* 적으로
* 적인
* 져야
* 지만
* 하고자
* 하기
* 하는
* 하되
* 하려
* 하면
* 하여
* 하지
* 한
* 할
* 해
* 해서
* 해야
* 했기
* 했던
* 히

註 : * 좌우절단표시.
% 조사절단표시.

<附錄 2> 助 詞 리 스 트

的
的으로
下에
가
과
과는
과의
나
는
도
된
들
들도
들에
들에게
들은
들을
들의
들이
라는
란
로
로부터
로서
로서의
로써
를
마다
만큼
보다
보다는
부터
부터는
에
에게
에게는
에는
에도
에서
에서는
에서도
에서의
에의
와
와는
와의
으로
으로는
으로부터
으로서
으로서의
은
을
의
이
이나
이라
이라고
이라는
이란
적
적으로
처럼
하에

<附錄 3>　　實驗 데이터

〈데이터 1〉

企業年金會計에 관한 硏究

企業年金會計의 構造를 고용주의 會計와 고용주 이외의 獨立된 報告實體인 年金制度의 會計로 나눈다면, 고용주 會計시스템의 會計處理는 年金費用이 發生할 때에 借邊에 年金費用, 貸邊에 年金負債로, 年金基金을 積立할 때에는 借邊에 現金으로 會計處理된다. 年金負債나 先給年金資産의 크기의 기초가 되는 것은 年金費用이다. 왜냐하면 年金費用의 크기가 먼저 결정되고, 그 다음에 拂入金額의 크기에 따라 年金負債의 先給年金資産이 年金負債 때문에 發生한 利子費用, 3. 企業年金制度의 給與條項의 修正이나 變更으로 인하여 발생한 過去勤務原價의 償却額, 4. 年金制度가 보유하고 있는 年金基金資産의 投資收益, 5. 保險統計的 假定과 實際의 差異로 방생한 利得과 損失의 償却額과 같은 이상의 5가지 合計로 구성된다. 이 5개의 구성요소가 年金費用이라는 단일의 金額으로 통합되어 報告되고, 이 年金費用의 크기가 타 計定의 크기에 영향을 미치고 있다. 企業年金會計에서는 企業年金制度의 利害當事者들의 經濟的 意思決定에 有用한 會計情報가 고용주의 財務諸表와 年金制度의 財務諸表에 公示되어야 한다.

〈데이터 2〉

韓·日貿易의 國際分業패턴에 관한 硏究－理論的 規範論에 입각한 對日貿易逆調是正을 위한 接近

對日貿易逆調라는 貿易不均衡問題는 지난 20여년간 다방면에 걸쳐 韓國經濟의 끊임없는 論題가 되었다. 그러나 한편으로 그 동안의 硏究가 對日貿易逆調是正의 當爲性에 대한 연구의 중요성을 인식하지 못하였고, 또한 전체적으로 새로운 視角에서 기존의 연구를 재분석 검토 종합하는 批判能力이 부족하였기 때문에, 그리고 다른 한편으로는－ 결국은 이상과 같은 認識不足과 批判能力의 不足으로 초래되는 것이지만 －對日貿易逆調問題에 대한 認識과 是正方法에 있어서도 韓·日間에 상당한 차이를 보여왔기 때문에 對日貿易逆調의 是正에 큰 어려움이 있었다. 본 硏究는 이러한 問題를 인식함으로써 對日貿易逆調를 새로운 視角에서 재음미하고 기존의 연구와는 다른 해석을 내리려는 시도라고 할 수 있다. 이를 위하여 理論的 規範論이라는 패러다임을 제시하였다. 이 패러다임은 對日貿易逆調問題를 4가지의 硏究課題(1. 對日貿易逆調의 發生原因, 2. 對日貿易逆調是正의 當爲性, 3. 對日貿易逆調의 是正方法, 4. 對日貿易逆調의 是正되지 않는 理由)로 구체화시키고 설정함으로써 피드백 과정을 통하여 나머지 세가지 硏究課題를 再分析 檢討 綜合하는 硏究方法이다.

<附錄 4>　　키워드 抽出結果

〈데이터 1〉

C 기업연금회계+에
D 기업
D 연금
D 회계
D 기업연금
D 연금회계
C 기업연금회계+의
D 기업
D 연금
D 회계
D 기업연금
D 연금회계
C 구조+를
K 고용주+의
C 회계+와
K 고용주
K 이외+의
C 보고실체+인
D 보고
D 실체
C 연금제도+의
D 연금
D 제도
D 연금제
C 회계+로
K 고용주
C 회계+시스템의
C 회계처리+는
D 회계
D 처리
C 연금비용+이
D 연금
D 비용
C 차변+에
C 연금비용
D 연금
D 비용
C 대변+에
C 연금부채+로
D 연금
D 부채
C 연금기금+을
D 연금
D 기금
C 차변+에
C 현금+으로
C 연금부채+이나
D 연금
D 부채
C 선급연금자산+의
D 연금
D 자산
C 연금비용+의
D 연금
D 비용
K 그
C 불입금액+의
D 금액
C 연금부채+의
D 연금
D 부채
C 선급연금자산+이
D 연금
D 자산
C 연금부채
D 연금
D 부채
C 이자비용
D 비용
C 기업연금제도+의
D 기업
D 연금
D 제도
D 기업연금
D 연금제도
C 급여조항+의
D 급여
D 조항
C 수정+이나
C 변경+으로
C 과거근무원가+의
D 과거
D 근무
D 원가
C 상각액
C 연금제도+가
D 연금
D 제도
D 연금제
C 연금기금자산+의
D 연금
D 기금
D 자산
D 연금기금
C 투자수익
D 투자
D 수익
C 보험통계+的
D 보험
D 통계
C 가정+과
C 차이+로
C 이득+과
C 손실+의
C 상각액+과
K 이상+의
K 가지
C 합계+로
K 개+의
K 구성요소+가
D 구성
D 요소
C 연금비용+이라는
D 연금
D 비용
K 단일+의
C 금액+으로
C 연금비용+의
D 연금
D 비용
C 타
C 계정+의
C 기업연금회계+에서는
D 기업
D 연금
D 회계
D 기업연금
D 연금회계
C 기업연금제도+의
D 기업
D 연금
D 제도
D기업연금
D 연금제도
C 이해당사자+들의
D 이해
D 당사자
C 경제+的
C 의사결정+에
D 의사
C 회계정보+가
D 회계
D 정보
K 고용주+의
C 재무제표+와
D 재무
C 연금제도+의
D 연금
D 제도
D 연금제
C 재무제표+에
D 재무

註 : K : 한글　　J : 일본어　　C : 한자어　　D : 분리된 단어　　A : 영숫자

〈데이터 2〉

C 한일무역+의
D 한일
D 무역
C 국제분업+패턴에
D 국제
C 규범런+에
C 대일무역역조시정+을
C 대일무역역조+라는
D 대일
D 무역
D 대일무역
D 무역역조
C 무역불균형문제+는
K 지난
A 20
K 여년간
K 다방면+에
C 한국경제+의
D 한국
D 경제
C 논제+가
K 한편+으로
K 그
C 대일무역역조시정+의
C 당위성+에
C 시각+에서
K 기존+의
C 재분석
C 비판능력+이
D 비판
D 능력
K 부족하였기
D 부족
K 한편+으로는
K 이상+과
C 인식부족+과
D 인식
D 부족
C 비판능력+의
D 비판
D 능력
C 부족+으로
C 대일무역역조문제+에
C 인식+과
C 시정방법+에
D 시정
C 한일간+에
K 차이를+를
C 대일무역역조+의
D 대일
D 무역
D 대일무역
D 무역역조
C 시정+에
K 어려움+이
C 대일무역역조+를
D 대일
D 무역
D 대일무역
D 무역역조
C 시각+에서
K 기존+의
K 해석+을
C 규범론+이라는
K 패러다임+을
K 패러다임+은
C 대일무역역조문제+를
K 가지+의
C 연구과제
D 과제
C 대일무역역조+의
D 대일
D 무역
D 대일무역
D 무역역조
C 무역
D 대일
D 무역
C 발생원인
D 발생
D 원인
C 대일무역역조시정+의
C 당위성
C 대일무역역조+의
D 대일
D 무역
D 대일무역
D 무역역조
C 시정방법
D 시정
C 대일무역역조+가
D 대일
D 무역
D 대일무역
D 무역역조
C 이유
K 이중
K 번
C 연구과제+를
D 과제
C 규범론+의
C 분석시각+으로
D 시각
D 분석시
K 피드백과정+을
D 피드백
K 나머지
K 세가지
C 연구과제+를
D 과제
C 재분석