중국어 정보검색을 위한 확장된 바이그램 분할기법

김운[O] 강지훈 맹성현
충남대학교 컴퓨터과학과
한국정보통신대학교
{wkim[O], jhkang}@cs.cnu.ar.kr,
myaeng@icu.ac.kr

An Extended Bigram Segmentation Method for Chinese Information Retrieval

Yun Jin[O] JiHoon Kang SungHyon Myaeng
Dept. of Computer Science, Chungnam National University, Korea
School of Engineering, Information and Communications University, Korea

요 약

중국어 문장은 영어와 한국어와 달리 단어에 대한 명확한 한계가 없기 때문에 중국어 정보검색 시스템에서는 중국어 문장에 대한 색인 작업을 각각의 글자를 기본단위로 자르는 방법을 사용하거나 또는 단어의 한계에 관한 정보가 이미 제공된 단어 사전을 이용하여 색인하는 방법을 사용하고 있다. 하지만 이 두 가지 방법은 모두 장단점이 있다. 본 논문에서는 이 두 가지 방법의 장점을 취하고 단점을 보안하는 방법으로 확장한 바이그램 분할기법을 제안하려 한다. 이 방법은 실용성이 있으며, 검색성능 향상을 도모하였다.

1. 서 론

중국어 정보검색 시스템은 중국어 문장에 대한 색인 작업을 영어나 한국어와는 다르다. 그 주요한 원인은 중국어 문장에는 단어에 대한 명확한 한계가 없다. 이런 중국어 문장 분할기법은 보다 깊은 중국어 문장 분석을 필요로 하게 된다.

중국 중국어 문장 분할기법은 크게 두 가지로 나뉘는데 한 가지는 단어 사전을 기반으로 한 방법이고 다른 한 가지는 각각의 글자의 개수를 기반으로 한 방법이다[1,2]. 그 중 단어 사전을 기반으로 한 방법은 미리 만들어놓은 단어의 한계에 관한 정보를 필요로 하고 글자를 기반으로 한 방법은 문장 중의 연속된 글자들을 자르는 기본 단위를 필요로 한다[3].

단어 사전 기반 분할방법은 단어 사전에 있는 단어는 분할을 하지만 사전에 없는 새로운 용어, 고유명사 등과 같은 단어는 한 글자로 잘리는 단점을 갖고 있어 실제 응용도 적다. 반면, 글자 개수를 기반으로 한 분할방법은 언어적 자원을 필요로 하지 않고, 구현이 쉽고, 특히 바이그램 글자기반의 방법은 검색 성능이 우수하다는 장점이 있으나[4] 역-색인 파일이 크고 검색하는데 보다 많은 시간이 걸린다는 단점을 갖고 있다.

본 논문에서는 위의 두 가지 분할방법의 장점을 취하고 단점을 보안하는 방법으로 확장된 바이그램(Bigram)방법을 제안한다.

2. 관련 연구

2.1 사전 기반의 분할 방법

중국어 정보검색에서 사전 기반의 방법 중 최장일치(Maximum Matching)알고리즘은 가장 대표적이다[5]. 최장일치알고리즘은 중국어 문장의 첫 글자부터 시작하여 그 뒤에 나오는 글자 리스트를 미리 만들어 놓은 단어 사전과 비교하면서 단어 사전 중에 이 첫 글자로부터 시작되는 가장 긴 단어가 있는지를 찾는 알고리즘이다. 이런 최장일치 알고리즘을 적용하여 발견된 단어만큼 중국어 문장을 자른 후 그 다음 글자부터 시작하여 이 과정을 반복해 나간다.

최장일치 방법은 단어 사전을 의거하기 때문에 분할된 색인 용어에 실제 단어가 나올 확률이 높으며 정확하게 분할될 확률도 높다는 장점이 있다. 그리하여 이 방법은 중국어 기계 번역 시스템이나 중국어 구문분석 시스템에 많이 사용되고 있다. 그러나 이 방법은 단어 사전에 의거하기 때문에 단어 사전의 영향을 많이 받는다. 특히 날마다 새롭게 산출되는 용어를 정확히 분할하기 위해 단어 사전을 확장해야 하는데 이렇게 하려면 많은 인력과 비용이 필요하며[3] 빈도수가 낮은 고유명사, 외래어들은 사전에 등록되어 있지 않다. 이런 미등록 용어들은 분할 시에 한 글자씩 잘리는 단점이 있다.

2.2 글자 기반의 분할 방법

글자 기반의 분할방법은 중국어 문장을 무조건 글자의 개수를 기본단위로 자르는 방법이다. 그 중 대표적인 방법은 바이그램방법이다. 바이그램방법은 중국어 문장을 두 글자를 기본단위로 자르는 방법으로써, 그 기본원리는 대부분의 중국어 단어가 두 글자로 이루어졌으며, 사용빈도 또한 기타 단어에 비해 많은 중국어의 언어적 특성을 이용한 것이다[3,6]. 따라서 바이그램기법에 의해 생성된 용어들 중 실제 두 글자 단어가 출현할 확률이 높고[7], 실제 단어가 아니더라도 질의를 같은 바이그램방법으로 분할하여 생성된다면 정보검색 목적을 달성할 수 있다.

바이그램방법은 글자를 기본단위로 자르기 때문에 아무런 언어적 자원도 필요로 하지 않고 구현이 쉬우며 결과도 우수하다[1,4]. 바이그램방법은 중국어 문장분석 기법이긴 하지만 정보검색에만 사용될 수 있고 기계번역이나 자연어처리에는 부적합하며, 정확하지 않은 용어들이 많이 산출되어 역-색인 파일이 크며, 검색 시 보다 많은 시간이 소요된다는 단점이 있다.

3. 확장된 바이그램 분할 방법

본 논문에서 제안하고자 하는 확장된 바이그램 분할방법의 기본 아이디어는 분할 시 단어사전과 바이그램 방법을 동시에 사용하여 분할한다. 즉 중국어 문장을 바이그램기법으로 자르되 단어사전에 존재하는 연속된 글자리스트는 바이그램방식이 아닌 사전 기반의 기법으로 분할한다. 예를 들면 " 情報檢索" , " 結果分析" 등 글자리스트는 단어 사전에 실제 존재하는 연속된 두 개 단어들이다. 이런 단어 리스트들을 바이그램 방식인 " ' 情報' ' 報檢' ' 檢索' " , " ' 結果' ' 果分' ' 分析' " 등으로 자르지 않고 두 글자 단어사전을 중심으로" ' 情報' ' 檢索' " , " ' 結果' ' 分析' " 등으로 잘리게 함으로써 정확하지 않은 용어 " 報檢" , " 果分" 의 추가 생성을 줄일 수 있다.

확장된 바이그램방법의 장점은 정확하지 않은 용어 추가 생성을 줄임으로써 바이그램방식에서의 역-색인파일 크기가 큰 단점을 보안할 수 있으며, 단어사전 중 존재하지 않는 고유명사, 외래어, 새로운 용어들은 바이그램방식으로 잘리게 함으로써 단어사전 분할방법에서의 한 글자로 잘리는 단점을 보안할 수 있다. 또한 바이그램방식과 단어사전방식을 결합함으로써 이 두 가지 방법의 장점을 효과적으로 살릴 수 있다.

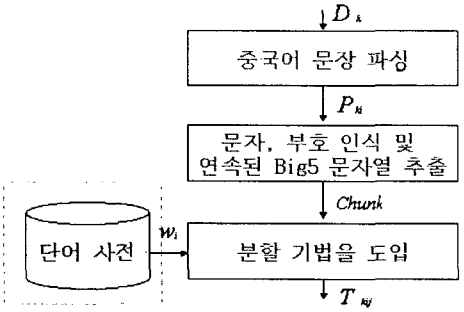
확장된 바이그램방법의 알고리즘은 중국어 문장의 첫 글자부터 시작하여 바이그램방식으로 자르면서 이렇게 잘린 용어들을 실제 단어인지를 두 글자 사전과 비교한다. 비교를 통하여 연속된 두 글자 단어가 출현하면 앞의 단어만 정확한 단어로 보고 분할한 후 뒤의 단어는 그 다음 글자 리스트 중 연속된 단어가 뒤따르는지를 비교한다.

본 논문에서는 색인을 더욱 효과적으로 하기 위해 위의 예제에서 설명한 " 2+2" 패턴 외에도 " 2+1+2" 등 패턴도 함께 처리하였다. 이런 패턴들도 중국어 문장에서 출현빈도가 많은바 " 2+1+2" 의 예로는 " 科學的發展" 이며, 이때 " 的" 는 불용어이다.

4. 시스템 구현

4.1 중국어 문장 분할부분

중국어 문장 분할의 전체흐름도는 다음의 [그림 1]과 같다.



[그림 1 중국어 문장 분할흐름도]

중국어 문장 분할은 중국어 문서를 구절단위로 파싱하여, 이 구절 중 연속된 글자리스트를 추출하여 분할기법을 도입한다. 이런 과정을 거쳐 최종 용어가 추출된다.

4.2 검색기 부분

본 논문에서 사용한 검색기 모델은 2-Possion모델이다. 이 모델의 수식은 다음과 같다.

$$W = \sum_{i=1}^{n} (w_{di} \times w_{qi})$$

$$w_{di} = \frac{tf_{di}}{k_1\left[(1-b)+b\times\frac{document\ \ length}{average\ \ document\ \ length}\right]+tf_{di}} \times \log\frac{N-n+0.}{n+0.5}$$

$$w_{ti} = tf_{qi}$$

위 식에서 N은 문서 총수이고 n은 질의의 용어가 존재하는 문서의 수이다. 그리고 위 식에서 k_1은 2.0을 취하였으며 b는 0.75를 취하였다.

5. 실험 및 결과 분석

5.1 실험 데이터

본 논문에서 제안하고자 하는 확장된 바이그램방법의 검색 성능을 평가하기 위하여 NTCIR-3 [8] 중국어 컬렉션과 중국어 질의를 실험 데이터로 사용하였다. NTCIR-3의 중국어 컬렉션은 98~99년도 대만 뉴스이며, 컬렉션의 총 문서 개수는 381,681개이며, 컬렉션의 크기는 약 550MB로서 문서 당 평균 크기는 약 1.4KB이다.

본 실험에서 사용한 적합성 판정은 NTCIR-3 중의 Relaxed 판정 방법이다[8].

또한, 본 논문에서 사용한 중국어 단어 사전에는 중국어 단어가 모두 119,804개 있으며 그 중 두 글자 단어 수는

72,966개로서 가장 많으며 전체 단어 수의 60.9%을 차지한다.

5.2 실험

실험은 본 논문에서 제안한 방법의 유용성을 비교하기 위하여 위에서 설명한 바이그램방법, 최장일치방법 그리고 확장된 바이그램방법 등 순으로 나누어 하였다. 확장된 바이그램방법 실험에서의 " 2+1+2" 패턴 중에 사용한 한 글자 불용어는 36개로써, 사람이 인터넷 중국어 문서를 직접 읽고 모은 것이다.

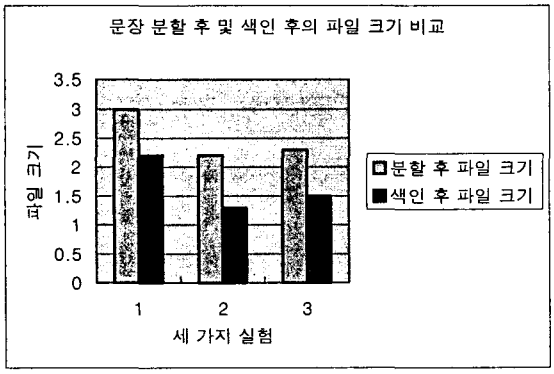
5.3 결과 분석

전체적인 실험 결과를 살펴보면 [표 1]과 같다. [표 1]에서 알 수 있듯이 바이그램 방법과 확장된 바이그램 방법은 모두 최장일치방법에 비해 검색 결과가 우수하다. 이것은 최장일치방법이 단어사전의 영향을 많이 받음을 알 수 있다. 따라서, 중국어 정보검색 시스템에 있어서 단어사전의 사용이 매우 큰 역할을 하며 잘못 사용하면 오히려 안 사용하는 것보다 못함을 알 수 있다.

[표 1 세 가지 분할기법 검색 결과표]

분할기법	Bigram	최장일치	확장된 Bigram
Avg. Pre.	0.3329	0.301	0.3409

[표 1]에서 알 수 있듯이 본 논문에서 제안한 확장된 바이그램방법은 순수한 바이그램 방법에 비해 그 검색 성능이 조금 우수하다.



[그림 2 분할 후 및 색인 후 파일 크기 비교도]

또한 본 논문에서 제안한 방법의 분할 후 파일 크기가 순수한 바이그램방법의 분할 후 파일 크기와 역-색인 파일이 모두 0.7GB 정도 작다[그림 2]. 이러한 파일 크기의 차이는 직접적으로 검색하는데 걸리는 시간과 연관된다.

6. 결론 및 향후 연구

본 논문에서는 중국어 정보검색에 있어서, 바이그램 방식에 비해 보다 정확하게 분할하는 단어사전 기반의 최장일치방법이 새로운 용어, 고유명사에 대한 치명적인 단점으로 인해 실제 검색시스템에 도입하기엔 부적합한 점과, 바이그램방법의 단순한 점에 비해 검색성능이 우수하며, 실제 검색기에 많이 도입되고는 있지만 역-색인파일의 크기가 많아 검색하는데 보다 많은 시간을 소요한다는 점을 감안하여 확장된 바이그램방법을 제안하였다.

실험을 통하여 확장된 바이그램방법은 검색성능이 위의 두 가지 방법보다 우수하며, 또한 분할 후와 색인 후의 파일 크기가 작아 검색하는데 보다 적은 시간이 소요됨을 알 수 있다.

향후 연구로는 본 논문에서 제안한 확장된 바이그램방법의 검색성능을 보다 향상시킬 수 있는 방법을 모색하려 한다.

참고문헌

[1] J.Nie and F.Ren. ' Chinese Information retrieval : Using characters or words ?' , In *Information Processing and Management*, 35 : 443-462, 1999.

[2] X.Huang and S.Roberton. ' A Probabilistic Approach to Chinese Information Retrieval, Theory and Experiments' . In Proceedings of the BCS-IRSG 2000

[3] F.Peng, X.Huang, D.Schuurmans, N,Cercone, S.E.Robertson. ' Using Self-Supervised Word Segmentation in Chinese Information Retrieval' , ACM SIGIR 2002.

[4] A. Chen, J. He, L. Xu, F. C. Gey, J. Meggs, ' Chinese Text Retrieval Without Using a Dictionary' , ACM SIGIR, Pages 42-49, 1997.

[5] Chen Keh-jian and Shing-Huan Liu, ' Word identification for Mandarin Chinese sentences' , In Proceedings of COLING-92, Pages 101-107, 1992.

[6] FDMC. Xiandai hanzi pinlu cidian (Frequency dictionary of modern Chinese). Beijing Language Institute Press, 1986.

[7] Y.R Chao, ' A Grammar of Spoken Chinse' , University of California Press, Berkeley, 1968.

[8] NTCIR. http://research.nii.ac.jp/ntcir/