2001년 한국정보처리학회 춘계 학술발표논문집 제8권 제1호

게임 정보검색을 위한 자동색인 및 신조어 처리 시스템 구현

이상준, 류근호
충북대학교 데이터베이스 연구실
{sjl, khryu }@dblab.chungbuk.ac.kr

Implementation of the Automatic Indexing and New Term Processing System for Game Information Retrieval

Sang-Joon Lee, Keun Ho Ryu
Database Laboratory, Chungbuk National University

요약

오늘날 국내외에 인터넷 보급의 대중화가 점차 확대되고 네트워크을 이용하는 게임의 증가에 따라 게임에 관련된 웹 문서에 대한 사용자의 요구가 증가되고 있다. 기존의 수작업에 의한 색인 방식은 많은 전문인력, 시간, 경비 등을 필요로 하기 때문에, 기하급수적으로 증가하는 웹 상의 정보를 처리하기에는 이미 그 한계에 이른 실정이다. 이러한 문제점의 해결을 위해 컴퓨터를 이용한 자동색인 시스템의 개발은 매우 중요하고 시급하다.

더구나 게임 분야에서 있어 신조어는 너무나 급속히 생성되고 있다. 따라서 이러한 신조어 처리는 효과적인 자동색인을 위한 중요한 요소이다. 이 논문에서는 사용자들에게 보다 적합하고 안정적인 게임 정보를 제공하기 위해 게임 용어 사전을 이용한 자동색인과 신조어 처리 시스템을 설계, 구현한다. 자동색인 및 신조어 처리를 위해 게임용어사전, TF-IDF, n-gram 추출법을 이용한다.

1.서론

오늘날 인터넷의 대중화는 정보의 양적팽창, 주제의 세분화, 정보원의 확산화 등을 의미한다. 인터넷의 발달과 더불어 네트워크 게임의 증가는 게임에 관한 웹 문서에 대한 사용자의 요구로 나타나게 되었다. 한편, 과거로부터 진행되어 온 수작업에 의한 색인 방식은 많은 인력, 시간, 경비 등을 필요로 하므로 정보량이 급증함에 따라 이미 한계에 이른 실정이다. 이러한 문제점은 자동색인과 새로 발생하는 용어를 처리하는 새로운 색인에 누적시킴으로서 해결할 수 있다. 따라서 이 연구에서는 전문적인 게임용어 사전을 이용한 색인추출과 보다 나은 색인효율을 높이기 위해서 게임명사사전과 n-gram 기법을 이용하여 게임용어를 추출한다. 또한 게임이라는 세분화된 분야에
있어, 신조어의 기하급수적인 증가는 자동색인에
있어서 치명적인 오류를 범할수 있으므로 이러한 많은 신조어의 추출에 대한 방안을 제시, 구현한다. 이 논문의 구성은 2장에서는 자동색인 기법에 대해서 살펴보고, 3장에서는 게임용어 사전과 n-gram 기법을 이용한 게임용어 자동색인 및 신조어 처리 시스템을 설계 및 구현하고, 5장에서 구현된 자동색인 및 신조어 처리 시스템의 성능을 평가 한다. 마지막으로 6장에서는 결론에 대해서 기술하였다.

2.관련연구

이장에서는 자동색인을 위한 기법으로써 TF(Term Frequency)-IDF(Inverted Document Frequency), 언어학적 분석 기법 그리고 n-gram에 대해서 고찰 한다.

2.1 TF-IDF

대부분의 웹 정보 검색 시스템은 사람이 직접 색인어를 추출하는 것이 아닌 프로그램에 의한 자동 색인 방법을 사용한다. 또한, 색인된 문서들간의 우선순위를 결정해야하는데, 우선순위를 결정하는 데 가장 많이 사용되는 것은 TF-IDF 알고리즘이다. TF는 한 단어가 한 문

서내에 등장하는 횟수를 나타내고 DF(Document Frequency)는 한 단어가 검색된 N개의 문서의 집합 중에서 몇 개 문서에 등장하는가를 나타낸다. 특정 검색어가 한 문서에서 많이 나타난다면 그 문서는 해당 검색어에 대해 중요한 문서라고 판단할 수 있지만 여러 문서에 걸쳐 모두 나타난다면 그 단어에 대한 중요도는 떨어진다고 볼 수 있다. 따라서 문서의 우선 순위를 구하려면 TF값과 DF의 역인 IDF값을 곱한 값으로 나타낸다 [1, 5, 7, 8].

2.2 언어학적 분석 기법

언어학적 분석 기법은 형태소 분석, 구문 분석, 의미 분석을 이용하여 색인어를 선정하는 방법으로 구문 분석과 의미 분석을 이용한 색인어 생성기의 구성은 아직은 매우 어렵고, 형태소 해석을 근거로 색인어를 선정하는 방법은 많이 시도되고 있으며, 현재 색인어 추출의 정확성과, 속도 향상을 위하여 연구하고 있다. 형태소 분석을 이용한 방법으로는 불용어 제거 기법이 있는데 이 기법은 텍스트 내의 각 단어를 분리한 다음 불용어, 전치사, 조사, 관사 등과 고빈도 단어를 제외한 모든 단어를 색인어로 선택하는 기법이며, 단서 기법은 "요약", "제목", "입증하다" 등 주제를 축약적으로 표현해 주는 특정한 의미의 단어를 찾아 그 단어가 출현한 문장에서 나타나는 단어들을 색인어로 선택하는 방법이다[2, 6, 8].

3. 시스템 설계 및 구현

3.1 자동색인

이 장에서는 자동색인 시스템을 구성하고 있는 부분의 설계 및 구현방법에 대해 살펴본다. 자동생성기 시스템은 색인어 어절단위 추출 모듈, n-gram 색인추출모듈과 저장 모듈로 이루어져 있다. 어절추출 모듈은 게임용어 사전을 이용하여 게임 웹 문서에서 색인어를 추출한다. n-gram모듈은 전단계에서 색인어로 인정치 않은 색인어를 다시한번 용어사전을 비교하여 색인어를 추출하고 저장 모듈은 추출된 색인어와 문서의 정보를 데이터베이스에 저장한다.

표 1. 게임 용어 사전

한글 (키워드1)	영문 (키워드2)	의미 (설명)
스트렝스	Strength	힘을 나타낸다. 들고 다닐 수 있는 무게, 맨손, 칼 공격의 데미지에 영향을 준다....
투시력	Clairvoyance	사정거리:특별, 지속시간:즉시, 캐스팅 시간:3, 효과범위:특별, 마법방어력계산:없음. 마법사는 ...
아이스 블링거	I c e BlIner	얼음속성의 세검으로 공격력 40, 한계력 100...

게임 색인어 생성기에서 사용하는 게임 용어 사전의 구조는 표1과 같다.

그림1은 색인어 생성기 수행 순서도이다.

일단 웹 문서가 들어오면, 필요한 부분의 태그를 제외한 나머지를 색인과정에 집어넣게 된다. 일단 어절단위의 색인을 실시하게 되며 이 과정에 게임용어사전에 비교하여 용어가 있으면 바로 게임정보로 인식하여 게임 정보 DB에 넣게되며, 없다면 다시한번 n-gram 기반 추출법을 이용해 모든 경우의 수에 대하여 색인을 추출하여 다시 한번 게임용어 사전과 비교하여 존재유무를 파악한다. 만약 이 경우에도 나타나지 않는 색인어는 후보 신조어DB에 집어넣는다[3].

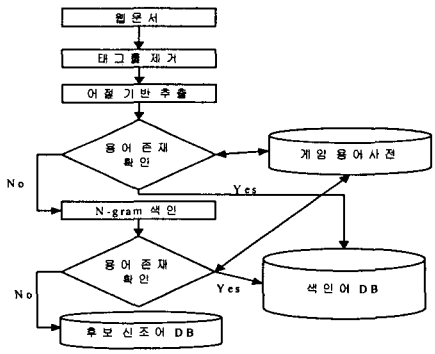

그림 1. 자동 색인어 생성 수행도

이 두 방법을 사용한 이유는 포괄적인 분야가 아닌 게임이라는 특정 영역에 주제가 한정되어 있는 시스템에서 최종 사용자가 전문용어에 익숙해 있는 경우 매우 효과적이기 때문이다. n-gram 추출기법으로 사용자 질의의 잘못된 띄어쓰기나, 복합명사와 같은 어절단위추출에서 불가능한 색인추출에 대해 정확한 결과를 얻기 위해 사용되었다.

일단 한 문자로 된 문자는 어절단위에서 게임용어 사전에 있는 것만 색인어로 인정하며 n-gram색인 추출시에는 색인어로써 제외시켰다. 게임 같은 경우 특수하게 한 글자로 된 문자가 많다. 예를 들면 "활","검","칼" 같은 용어가 많기 때문이다. 따라서 본 논문에서는 한 글자 용어는 어절단위에서 걸러내며 n-gram 추출시에는 한 글자는 색인어로써 가치를 인정하지 않으며 기본적으로 두 글자 이상으로 된 것을 색인 할 가치가 있는 것으로 보았다. 따라서 "항공모함" 이라는 용어가 있을 때 n-Gram을 이용할 경우는 "항공","공모","모함","항공모","공모함","항공모함"이라는 색인어를 검출할수 있다 [4].

3.2.신조어 추출

이 논문에서 후보 신조어 중에서 신조어를 추출하기 위해 사용자 질의어를 이용하였다. 일반적인 방법으로

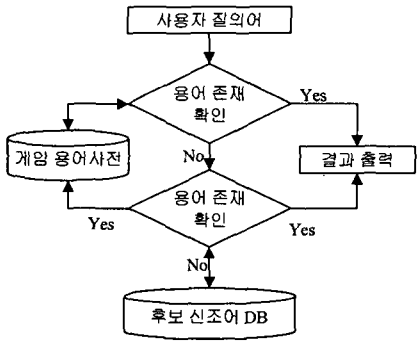

그림 2. 신조어 생성시 수행 순서도

사용자에게 몇가지 예를 주어 그중에서 선택하는 방법을 취하고 있지만, 이 방법에 있어 사용자의 무응답의 비율이 너무 높아 현실적인 한계가 있는 실정이다. 따라서 이 논문에서는 사용자가 게임검색엔진을 이용한다는 자체로 어느정도 게임용어에 익숙한 사용자로 가정하며, 일정수 이상의 사용자의 질의가 후보 신조어 데이터베이스(DB)에 있는 단어의 경우 이것을 신조어라 가정하고 실험을 하였다. 그림 2는 신조어 생성기 수행 순서도이다. 이 방법에 있어 저빈도 용어가 신조어로 처리되는 것을 방지하기 위해 경계값을 주어 그 경계값 이상의 단어만을 신조어로 선택하는 방법을 사용하였다. 그리고 후보 신조어 데이터베이스에 들어가는 용어들은 가중치를 부여받아 그 값으로 중요도를 결정한다.

각각의 가중치 W는 어절기반 추출의 경우 10이라는 값을 주었고 어미,조사를 제거 단어의 경우 5, N-gram을 거친 경우 가중치 1을 주었다. 가중치 W는 무작위 웹문서 2,000개를 선정하여 가중치를 결정하였다. 각각의 용어는 이러한 가중치 결정방법에 의해 중요도가 결정되며 고빈도 용어를 색인가능성 있다고 보았다. 따라서 저빈도 용어는 삭제함으로써 후보 신조어 데이터베이스의 크기를 고정시킨다. 또한 사용자의 질의에 있어서도 동일한 용어빈도수가 일정값 이상인 것만을 인정하였다.

4.실험 및 성능평가

이 논문에서 제안하는 게임 자동색인 기법의 성능을 평가하기 위해 웹 문서상의 게임관련문서들을 대상으로 실험을 수행하였다. 실험방법은 수작업으로 추출된 색인어와 이 논문에서 제안한 방법에 의해 추출된 색인어를 비교하여 색인어의 정확율을 측정하는 방법을 사용하였다. 먼저, 실험에 사용된 문헌의 특성을 살펴보면 표2와 같다.

표 2. 실험 대상 문헌 정보

문헌의 수	문헌의 크기	추출된 어절수
1000 개	20000 KByte	36,583 개

웹 문서에 대해 자동색인 기법 및 신조어 처리에 대한 결과는 표 3과 표 4와 같다. 여기서 사용자 질의는 2,345건 이었다.

표 3. 자동 색인기법에 대한 성능평가 결과

수작업 색인어의 수	7,862개
자동 추출된 색인어의 수	7426개
일치 색인어의 수	6984개
색인어의 정확률	88.83%

표 4. 신조어 추출에 대한 성능평가 결과

수작업의 신조어 수	436개
자동 추출된 신조어 수	189개
일치 신조어의 수	174개
신조어 정확률	39.90%

여기서 색인어의 정확률은 게임명사사전을 이용한 수치이며 만약 색인어가 사용자 오류로 인해 복합명사화 되었다면 해당 복합명사를 n-gram기법을 이용하여 각각의 단순명사를 별개의 색인어로 간주하였다.

표3에서 추출된 색인어의 정확율은 다음의 공식에 의해 계산되었다.[8]

$$\text{정확률} = \frac{\text{수작업색인어} \cap \text{자동추출색인어}}{\text{수작업색인어}}$$

표 5. 신조어 처리후 성능평가 결과

수작업의 색인어의 수	7,862개
자동 추출된 색인어의 수	7,570개
일치 색인어의 수	7,158개
색인어의 정확률	91.05%

이 논문에서 제안한 자동색인 기법을 사용함으로써 88.83%라는 결과를 얻었으며, 신조어 처리를 함으로써 단순히 사전만을 이용하는 결과보다 2.21% 정확률을 높였다. 물론 사용자의 많은 질의어가 들어온다면 신조어 처리에 있어 정확률이 더 높아질수 있다.

5.결론

이 논문에서는 게임 용어 사전과 n-gram 기법을 사용하여 웹 상에서 게임에 관한 정보를 효과적으로 색인 할 수 있는 자동색인 시스템을 설계, 구현하였다. 이를 위해 먼저 자동색인 기법들의 분석을 통하여 자동색인 기법들에 대해 살펴보았다. 이 논문에서 구현한 자동색인 및

신조어 처리시스템은 단순명사 뿐만 아니라 복합명사까지도 효율적으로 색인할 수 있다.

또한 후보 신조어 데이터베이스에 가중치를 주어 저빈도 용어의 신조어 처리되는 것을 방지하며, 이러한 신조어 처리를 함으로써 기존의 게임사전만을 이용했을 때 정확률 88.83%에서 91.05%로 전체정확율을 2.21% 올릴 수 있었다. 향후 연구방향으로는 신조어 처리에 있어서 정확률을 높이며 사용자 질의중에서 잘못된 질의를 검증할수 있는 방안에 대한 연구가 필요하다.

또한 이미지 연구에 관련하여 웹 문서에 포함되거나 연결된 이미지를 설명하는 텍스트를 이용하는 방법에 대한 연구가 필요하다.

참고문헌

[1] G. Salton and M.E. Lesk, "Computer Evaluation of Indexing and Text Processing", Reading in Infromation Retrieval, Morgan Kaufmann publishers ,1997

[2] L.B.Doyle, "Indexing and Abstracting by Association", Reading in Infromation Retrieval, Morgan Kaufmann publishers ,1997

[3] W.Bruce Croft, "Information Retrieval System: Theory and Implementation", Kluwer Academic Publishers, 1997

[4] Joon Ho Lee, Jeong Soo Ahn, "Using n-Grams for Korean Text Retrieval", SIGIR, pp.216-224, 1996

[5] Ricardo Baeza-Yates, Berthier Ribeiro-Neto,"Modern Information Retrieval", Addison-Wesley, 1999

[6] 류근호, 이재환 공역, "정보저장 및 검색", 시그마프레스, 2000

[7] 류근호, 김진호 공역, "정보검색", 시그마프레스, 1995

[8] 정영미, "정보검색론", 구미무역(주) 출판사, 1993