제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

주제어의 유사도 분석에 기반한 협업문서 생성제어 시스템

조성웅*, 원용관**, 이도헌***, 이귀상*
*전남대학교 컴퓨터정보학부
**전남대학교 정보통신공학부
***한국과학기술원 바이오시스템학과
e-mail:sucho@dbcore.chonnam.ac.kr

A Control System for Creation of Collaboration Documents based on Keyword Similarity Anaysis

Sung-Woong Cho*, Yong-Kwan Won**, Doheon Lee***, Guee-Sang Lee*
*School of Computer & Information, Chonnam National University
**School of Electronics & Computer Engineering, Chonnam National University
***Department of BioSystems, KAIST

요 약

인터넷은 공동작업과 지식정보 공유를 보다 쉽고 효율적으로 하기위해 시간과 공간을 초월하는 동적인 협업시스템이 필요하게 되었다. 새로운 형태의 지식공유 시스템인 위키(WIKI)는 연구원간의 자유스러운 정보교환을 보장하는 협업공간을 제공함으로써 지식정보의 생산성과 효율성을 극대화시킨다. 하지만 정보량이 방대해짐에 따라 공동의 주제를 가진 문서들이 중복되어 생성됨으로써 주제의 분산이 이루어져 정보공유의 힘을 약화시키는 문제점을 야기시킨다. 본 논문에서는 이러한 문제점을 해결하기위해 파서(parser), 문서 분류시스템, 유사성 측정시스템으로 구성된 협업문서 생성제어 시스템을 제안한다.

1. 서론

인터넷의 급속한 발전은 방대하고 널리 퍼져있는 정보를 집약시킴으로써 전세계를 하나로 묶어가고 있으며, 공동작업 공간으로서 지식의 공유와 배포를 웹상에서 가능하도록 하였다[1]. 인터넷상의 공동체 시스템인 위키(WIKI)는 효과적인 의사소통 및 지식공유 미디어로 사용자에게 편리하고 간편하게 사용할 수 있는 지식공유 공간을 제공함으로써 웹 기반의 협업시스템으로써 활용가치가 높아지고 있다.

위키시스템은 누구나, 아무 때나, 어디서건, 어느 것이든 자신의 의견을 삽입/수정/삭제할 수 있는 자유로움과 유연성을 제공함으로써 정보의 동적인 재개편이 쉽고 다양한 지식을 사용자가 자발적으로 재조직할 수 있다는 큰 장점을 가지고 있다. 하지만 자유로움과 유연성으로 인해 정보가 산재되어지고 동일한 주제에 대한 정보가 계속적으로 만들어져 주제의 분산이 이루어지는 단점을 가지고 있다.

본 논문에서는 동일한 주제문서에 대한 생성을 적절히 제어하는 협업문서 생성제어 시스템을 제안함으로써, 위키의 장점을 그대로 유지하면서 주제 분산의 단점을 보안했다. 제안하는 방법은 신규 문서 생성시 적절한 주제어를 도출하고 기존 문서 집합과의 유사도 평가를 통해 불필요한 별도의 협업문서 집단이 생기는 것을 방지함으로써, 주제의 분산을 줄이게 된다. 구현된 시스템은 한국과학기술정보연구원(KISTI)의 샘플 문서 150개를 이용하여 직접 관찰 결과와 시스템결과를 비교 실험평가 하였다[2].

2. 위키시스템

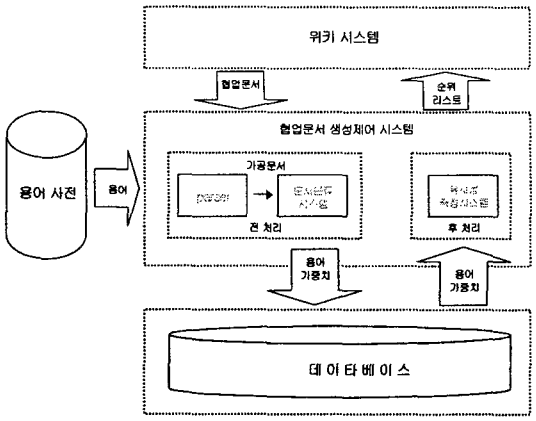
위키는 하와이어로 '빨리'라는 뜻으로 누구나 '자유롭게' 정보와 지식을 편집할 수 있는 지식공간을 제공하는 동적 프로그래밍이다. 초장기 위키는 다지

인 패턴을 연구하기 위해 개발한 자동화 도구였다. 이후 지속적으로 발전해, XP(eXtreme programming)라는 새로운 프로그래밍 패러다임으로 주제를 확장했고, 위키 자체와 그 외 다양한 정보를 담고 있는 '지식 그물'로 진화하였다. 위키의 근간을 이루는 엔진도 발전을 거듭해 펄, C, 자바, PHP, 파이션, 비주얼 베이직 등 대부분의 프로그래밍 언어로 구현된다. 또한 msql, mysql과 같은 데이터베이스도 지원한다[3][4].

누구나 자유와 참여로 이루어지는 위키의 힘은 다음과 같은 몇 가지 기능들에 의해 보여질 수 있다.
첫 번째, 누구나 자신이 원하는 주제에 알맞은 문서를 생성할 수 있는 공동저작기능. 즉, 모든 사용자가 시스템 어디든 문서이름만을 입력하면 문서를 추가할 수 있어 쉽게 자신이 언급하고자하는 주제의 문서를 추가할 수 있다.
두 번째, 누구나, 수시로 주제에 맞는 의견이나 결과를 추가, 삭제/수정할 수 있는 공동 편집기능.
세 번째, 기존의 문서에 내용을 수정하게 될 경우 자신이 쓰는 정보가 너무 길다고 느껴지면 새로운 문서를 만들어 연결하는 링크기능.
네 번째, 문서의 가시성을 위한 간단한 문서 formating 기능이 있다.
이런 기능들은 공동의 작업을 하는 사용자간에 정보를 공유하는 공간으로 아주 유용하게 이용되지만, 쉽게 문서를 생성하거나 수정하여 주제가 분산되는 문제점을 지니고 있다.

3. 시스템 설계 및 구현

제안하는 협업문서 생성제어 시스템은 특정 그룹에 속한 사용자가 어떤 주제에 대해 논의하고자 협업문서를 만들 때 중복된 문서가 있는지 보여주는 시스템이다.
(그림1)은 전체 시스템의 구조를 보이고 있다.
본 논문에서 제안하는 협업문서 생성제어시스템은 다음과 같이 3개의 단계로 구성된다.
1단계, 위키시스템 내의 협업문서들을 협업문서 생성제어 시스템에 가져와서 파서(parser)에 의해 tag, 공백문자, 접미사등 불용어를 제거한다.
2단계, 가공된 협업문서는 사전(시스템 내에서 논의되어질 주제에 관련되는 용어들을 모두 포함하고 있는 집합)의 용어들을 기준으로 문서들의 가중치값을 계산하여 데이터베이스에 저장한다.

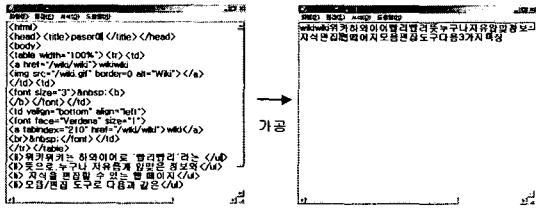


(그림1) 시스템 구성도

3단계, 사용자가 만들고자 하는 문서의 주제어 3개에 대해 데이터베이스에 저장된 협문문서들의 가중치값을 이용하여 각 문서의 순위값을 계산하고 문서 순위 리스트로 보여준다.

3.1 파서

문서 집합에서 너무 빈도가 높은 단어는 변별력이 좋지 않다. 어떤 문서집합에서 80%이상의 문서에 공통으로 출현한 단어의 경우 검색에 쓸모가 없는데 이러한 단어를 종종 불용어라 한다[5][6]. 불용어들은 문서 내에서 미리 걸러내 짐으로써 문서를 재구성시킬 수 있다. 따라서 불용어 제거는 불필요한 단어가 제거됨으로써 문서의 주제를 명확히 할 수 있을 뿐 아니라 문서의 크기를 작아지게 함으로써 색인 구조의 크기를 줄여주는 장점이 있다.



(그림2) 협업문서 가공 예

파서(parser)는 협업문서 내의 html태그, 접미사, 공백문자 등 문서의 주제를 파악하는데 필요 없는 불용어들을 제거한다. (그림2)는 파서를 이용하여 협업문서를 가공한 예를 보이고 있다.

3.2 문서분류시스템

용어에 대한 문서의 유사도는 문서d에 출현한 용

어 t의 빈도수를 측정함으로써 수치화된다. 이런 용어빈도수는 통상 TF요소라고 불리며 그 용어가 문헌의 내용을 얼마나 잘 표현하는가의 척도이다. 또한 용어에 대한 문서의 비유사도는 전체 문헌 컬렉션 중에서 용어 t가 출현한 문서 빈도수의 역수를 계산함으로써 구할 수 있는데 이는 IDF요소라고 불리며 IDF 요소의 사용의 동기는 많은 문서에 출현한 용어가 연관 문서와 비연관 문서의 구분이 쓸모가 없다는데 있다[6-8].

TF*IDF알고리즘을 이용한 문서들의 가중치값을 구하는 식은 아래와 같다.

$$W_{dt}\text{(가중치값)} = f_{dt} * w_t \qquad \text{식(1)}$$
$$w_t = \log(N/f_t)$$

f_{dt}: 문서 d에서 t인 용어의 출현 빈도수

w_t: 역문서 빈도수(N: 총 문서수; f_t: t인 용어의 출현한 문서 빈도수)

N: 총문서수

문서에 출현한 용어의 빈도수를 측정하기 위해 문자열 탐색 알고리즘을 이용하는데 본 시스템은 Brute Force 알고리즘을 사용하였다[9].

문서분류시스템은 사전에 존재하는 하나의 용어를 기준으로 각 문서들의 가중치 값을 적용하기 때문에 해당 용어가 다른 문서에 얼마나 자주 나타나는가를 적용시켜야 한다. TF*IDF 알고리즘은 문서의 가중치를 계산하는데 아주 좋은 알고리즘이지만 용어가 다른 문서에 얼마나 자주 나타나는가를 적용하지 못한다.

다음은 TF*IDF 알고리즘의 문제점을 해결하고, 문서간의 적절한 분류를 위해 본 논문에서는 제안하는 식이다.

$$f'_{dt} = freq_{d,t} / \sum_{d=1}^{N} freq_{d,t} \qquad \text{식(2)}$$

식(2)의 $f'_{d,t}$는 식(1)에서의 $f_{d,t}$를 알맞게 변형한 것으로 시스템내 모든 문서들에 용어t의 출연 빈도수를 적용한 것이다.

문서 \ 용어	용어1	용어2	용어3	....
문서1	0.34	0.07	0.43	...
문서2	0.36	0.45	0.03	..
..	..	..	..	..

(그림3) 가중치 테이블

문서분류시스템에서는 용어와 문서들간의 가중치를 계산하여 (그림3)과 같은 형식의 테이블을 데이터베이스에 만들어 정보를 저장한다.

3.4 유사성측정시스템

사용자가 만들고자 하는 신규협업문서에 해당하는 주제어 3개를 도출하여 시스템에 보내면 주제어들이 사전에 존재하는지를 확인한 후 문서분류시스템에서 저장해 놓은 가중치값을 이용하여 문서들의 순위값들을 계산하고 순위 리스트를 보여준다.

일반적으로 각 주제어에 대한 협업문서들의 가중치값 3개를 더함으로써 순위값을 계산할 수 있다[5]. 그러나 이와 같은 경우 사용자가 만들고자 하는 문서의 주제와 같은 주제를 가지는 문서가 다른 주제를 가지는 문서의 순위값보다 낮아지는 경우가 발생하는 문제점을 가지고 있다. 즉 3개의 주제어를 모두 포함하는 문서의 가중치값은 아주 특별히 하나의 주제에 대해 높은 가중치값을 가지는 문서보다도 더 낮은 순위가 된다는 것이다.

제안하는 유사도측정시스템에서는 이러한 문제점을 식(3)에서와 같이 문서가 주제어를 포함하는 확률값을 곱함으로써 해결하였다.

$$Rank(D) = (s/3) * \sum_{i=1}^{3} W_{d,i} \qquad \text{식(3)}$$

$w_{d,i}$: 주제어 i에 대한 문서 d의 가중치 값

s : 문서 d에서 포함하는 주제어 개수

4. 실험평가

본 논문은 제안하는 협업문서 생성제어 시스템의 신뢰성을 측정하기 위해 한국과학기술정보연구원(KISTI)의 샘플 문서 150개를 이용하여 실험평가 하였다[2]. 문서의 구조는 (그림4)과 같다.

```
#T
4분기 경기 계속 "호텔~/한은 2천4백사 조사
#S
실명제 실시로 국내경기가 3분기중 다시 급격한 하강곡선을 그렸으며,
4분기에도 투명한 경기회복을 기대하기 힘들 전망이다. 한국은행은 2일
전국 2천4백여개 업체를 대상으로 실시한 「기업경기조사」 결과 3분기
(7~9월)중 기업인의 체감경기를 나타내는 경기실사지수(BSI)가 당초
전망이나 2분기수준에 크게 못미치는 것으로 나타났다고 발표했다.
제조업의 경우, 실명제실시전 조사에선 3분기중 업황BSI가 1백9에 답해
본격적인 경기회복이 기대됐었으나, 실명제시행뒤인 지난 9월초 실시된
이번 조사에선 모 나와 당초전망보다 포인트나 큰 폭으로 줄어들었다.
이는 2분기중 보다도 포인트나 떨어진 수준이며, 그만큼 실명제실시후
기업의욕이 위축되어 있음을 반영하고 있다.BSI는 기업인이 현장에서
피부로 느끼는 체감경기를 나타내는 지수이며, 1백을 넘으면 「경기가
좋다」는 업체가 「좋지 않다」는 업체보다 더 많음을, 1백 이하이면 그
반대임을 나타낸다.한국은행은 또 4분기(~월)에도 제조업의 업황BSI
전망치가 1백에 미달하는 에 그쳐 경기실체 분위기가 여전할 것으로
예상했다.
#A
실명제 실시로 국내경기가 3분기중 다시 급격한 하강곡선을 그렸으며,
4분기에도 투명한 경기회복을 기대하기 힘들 전망이다.
#B
실명제 실시로 국내경기가 3분기중 다시 급격한 하강곡선을 그렸으며,
4분기에도 투명한 경기회복을 기대하기 힘들 전망이다. 한국은행은 2일
전국 2천4백여개 업체를 대상으로 실시한 「기업경기조사」 결과 3분기
(7~9월)중 기업인의 체감경기를 나타내는 경기실사지수(BSI)가 당초
전망이나 2분기수준에 크게 못미치는 것으로 나타났다고 발표했다.
#C
한국은행 조사 결과 4분기 경기회복 힘들 듯.
```

(그림4) 샘플 문서

(그림4)에서 #T는 문서의 주제를 나타내고, #S는 문서내용, #A는 문서내용의 10% 추출, #B는 문서내용의 30%추출, #C는 문서내용의 10% 요약을 나타낸다. 1번부터 150번까지의 문서내용이 중복되는 #A, #B, #C는 제거하고, 수작업으로 각 문서의 주제(#T)에 대해 수작업으로 용어 286개를 추출하여 사전을 구축하였다.

협업문서 생성제어 시스템을 실험하기 위해 직접관찰로 5개의 실험 표준을 만들어 (표1)와 같이 시스템 결과와 비교하였다. (표1)에서 <>안의 번호들은 1~150번까지의 샘플문서에서 주제어에 관련있는 문서의 번호가 오름차순으로 나열된 것이다.

(표1) 실험 비교

	주제어	직접관찰 결과	시스템 결과
1	북한,핵,안보리	<86, 1, 27, 58, 29>	<86, 27, 58, 1, 29>
2	러,엘친,모스코바	<148, 131, 30, 126, 84>	<131, 148, 30, 126, 84>
3	엘친,보수파,러	<131, 30, 148, 41, 126>	<126, 131, 30, 41, 148>
4	중국,핵,클린턴	<145, 3, 29, 121, 58>	<145, 4, 3, 27, 29>
5	경찰,검찰,골프장	<54, 47, 45, 39, 25>	<54, 45, 39, 47, 25>

5. 결론

본 논문은 신규 협업문서의 생성 시 적절한 제어를 통해 공동의 주제를 가지는 협업문서 생성을 제한하는 협업문서 생성제어 시스템을 제안하였다. 협업문서 생성제어 시스템은 문서 가공 역할을 하는 파서(parser), 사전의 용어에 대한 문서의 유사도를 측정하는 문서분류시스템, 신규 협업문서에 대한 주제어를 기준으로 기존 협업문서들의 순위를 계산하는 유사성 측정시스템으로 구성됐다. 결과적으로 사용자가 생성하고자하는 신규 협업문서에 대한 기존의 협업문서의 순위리스트를 보여줌으로써 중복된 주제의 협업문서가 생성되어지는 것을 막아, 시스템을 사용하는 사용자간의 원활한 정보공유와 지식창출을 효과적으로 할 수 있게 했다.

향후 시스템은 유사도 측정에 대한 용어와 용어간의 공기관계를 이용하여 결과도출에 좀 더 정확성을 부여하는 연구가 필요하다. 또한 동적 시스템을 위해 사전을 자동으로 구축하는 연구가 과제로 남아있다.

참고문헌

[]]Jay F. Nunamaker, jr. and Robert O. Briggs. Introduction to the Collaboration systems and Technology Track, proceedings of the 35th Hawaii International Conference on System Sciences, 2002.

[2] 김태희, 박혁로, 신중호, "검색/요약/필터링을 위한 텍스트 이해 모형 연구, 제3회 소프트과학 워크숍 논문집, 1999.

[3]http://www.zdnet.co.kr/anchordesk/todays/sypark/.

[4] http://lovol.net/ExtremeProgramming.

[5] 김명철외 5명, 최신정보 검색론, 홍릉과학출판사.

[6] W. B. Frakes and R. Baeza-Yates. Information Retrieval : Data Structures algorithms. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.

[7] Ian H. Witten, Alistair Moffat, Timothy C. Bell, Managing Gigabytes, VanNostrandReinhold.

[8] Gerard Salton and Christopher Buckley. Term weighting approaches in automatic text retrieval Information Processing & Management, Vol. 24, No. 5, pp. 513-523, 1988.USA, 1992.

[9] Cristian Charras, Tierry Lecroq, "Brute force algorithm", Handbook of Exact String-matching Algorithms, http://www-igm.univ-mlv.fr/~lecroq/string/index.html.