주제 분산의 억제를 위한 협업문서 생성제어 시스템

조성웅*[0] 원용관*** 이도헌** 이귀상*
*전남대학교 컴퓨터정보학부
***전남대학교 정보통신공학부
한국과학기술원 바이오시스템학과**
*sucho@dbcore.chonnam.ac.kr[0], { **ykwon, *gslee}@chonnam.ac.kr, **dhlee@mail.kaist.ac.kr

Collaboration Doument Ranking System for the Control of Subject dispersion

Sung-Woong Cho[0] 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

인터넷의 발전으로 단순한 co-browsing을 넘어선 다기능 협업시스템이 필요하게 되었다. 이러한 점에서 웹 저작 도구인 위키 시스템은 연구원들 간의 능동적이고 적극적인 정보교환을 위한 효과적인 시스템이다. 하지만 정보량이 증가함에 따라 공통된 주제의 문서가 다중 생성됨으로써 정보 공유의 힘이 분산되는 문제점을 발생시킨다. 본 논문에서는 파서(parser), 문서분류 시스템, 유사성측정 시스템으로 구성된 협업문서 생성제어 시스템을 제안한다. 결과적으로 협업문서 생성제어 시스템은 협업문서 생성을 제어함으로써 각 분야의 전문가들의 원활한 정보 공유와 지식창출을 효과적으로 할 수 있다.

1. 서 론

인터넷이 급속히 보급되고, 기능과 성능이 발전함에 따라, 초기의 정보 배포 및 공유 수단으로서의 역할은 물론, 여러 사람의 동적 공동 작업 공간으로서의 역할을 수행하게 되었다. 특히, 그러한 공동 작업을 위한 프레임워크로서 개발된 위키(WIKI) 시스템은 구조의 간편성과 사용의 편의성에 힘입어 협업 공간으로서 점점 활용도가 높아지고 있다.

위키시스템은 기본적으로 참여자 누구나 협업 문서를 임의로 생성하거나 삭제할 수 있도록 하는 유연성을 제공한다. 하지만, 그러한 유연성 때문에 동일한 주제에 대하여 하나 이상의 협업 문서가 생성될 수 있고, 결국 동일한 주제에 대한 정보가 여러 문서에 분산될 수 있다는 단점이 있다.

본 논문은 협업문서를 생성하는 방식을 적절히 제어함으로써, 위키 본래의 유연성을 유지하면서도 정보의 주제가 분산되는 것을 제한할 수 있는 구조를 제안한다. 제안하는 방법은 신규 문서 생성시 적절한 주제어를 도출하고 기존 문서 집합과의 유사도 평가를 통해 불필요한 별도의 협업 문서 집단이 생기는 것을 방지함으로써, 주제의 분산을 줄이게 된다.

2. 위키 시스템

위키는 하와이로 '빨리'라는 뜻으로 누구나 자유롭게 정보와 지식을 편집할 수 있는 지식공간으로 웹을 기반으로 한 동적 프로그래밍이다[1]. 위키는 문서 모음/편집 도구로 웹브라우저를 통하여 여러 사람이 공동으로 문서를 생성 및 편집할 수 있는 기능을 가지고 있다. 위키는 다음과 같은 대표적인 네 가지 기능을 가지고 있다.

첫 번째, 문서 추가 기능은 모든 사용자가 시스템 어느곳이든 문서이름만을 입력하면 문서를 추가할 수 있는 기능이다. 두 번째, 문서 수정기능은 모든 사용자가 시스템에 저장되어진 문서를 누구나 수정 할 수 있는 기능이다. 세 번째, 링크는 기존의 문서에 내용을 수정하게 될 경우 자신이 쓰는 정보가 너무 길다고 느껴지면 새로운 문서를 만들어 링크를 연결하는 기능이다. 네 번째, 문서의 가시성을 위한 간단한 text formating기능이 있다

[2].

이런 기능들은 공동의 작업을 하는 사용자간에 정보를 공유하는 공간으로 아주 유용하게 이용되지만, 쉽게 문서를 생성하거나 수정하여 주제가 분산되는 문제점을 지니고 있다.

3. 협업문서 생성제어 시스템

3.1 시스템 구조

제안하는 협업문서 생성제어 시스템은 사용하는 특정 그룹에 속한 사용자가 어떤 주제에 대해 논의하고자 협업문서를 만들 때 중복된 문서가 있는지 보여주는 시스템이다.

그림[1]은 협업문서 생성제어 시스템의 구조를 보여주고 있다.

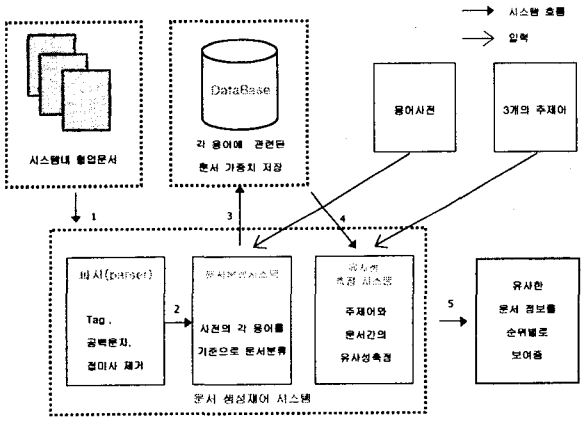

그림 1 협업문서 생성제어 시스템 구성도

시스템 전체흐름은 3단계로 구성되어진다.

1단계, 시스템내의 협업문서들을 파서(parser)에 의해 가져오게 되고 가져온 협업문서들은 tag, 공백문자, 접미사 등을 제거한다.

2단계, 가공된 협업문서는 용어사전 (시스템 내에서 논의되어질 모든 문서주제를 포함하는 용어들의 집합)의 용어들을 기준으로 문서들의 가중치값을 계산하여 데이터베이스에 저장한다.

3단계, 저장된 문서들의 가중치값을 가지고 사용자가 만들고자 하는 문서의 중복여부를 알아보기 위해 3개의 주제어에 대해 각 문서의 순위값을 계산하여 리스트로 보여준다.

3.2 파서(paser)

문서 집합에서 너무 빈도가 높은 단어는 변별력이 좋지 않다. 어떤 문헌집합에서 80%이상의 문헌에서 공통으로 출현한 단어의 경우 검색에 쓸모가 없는데 이러한 단어를 종종 불용어라 한다[3][4].

파서(parser)에서는 협업문서 내의 html태그, 접미사, 공백문자 등 문서의 주제를 파악하는데 필요 없는 불용어들을 제거한다.

3.3 문서분류 시스템

용어에 대한 문서의 유사도는 문서d에 출현한 용어 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[4][5][6].

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[7].

문서분류 시스템은 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, 문서들간의 적절한 분류를 위해 본 논문에서는 제안하는 식이다.

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

식(2)의 $f'_{d,t}$는 식(1)에서의 $f_{d,t}$를 알맞게 변형한 것으

로 시스템 내 모든 문서들에 용어t의 출연 빈도수를 적용한 것이다.

3.4 유사성측정 시스템

사용자가 만들고자 하는 문서에 해당하는 주제어 3개를 시스템에 보내면 이 주제어들이 사전에 존재하는지를 확인한 후 문서분류 시스템에서 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 가중치 값은 아주 특별히 하나의 주제에 대해 높은 가중치값을 가지는 문서보다도 더 낮은 순위가 된다는 것이다.

제안하는 유사도측정 시스템에서는 이러한 문제점을 문서가 주제어를 포함하는 확률값을 곱함으로써 해결하였다.

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

$w_{d,i}$: 주제어 i에 대한 문서 d의 가중치 값
s : 문서 d에서 포함하는 주제어 개수

4. 결론

시간과 공간의 제약을 초월하는 인터넷 협업 시스템의 활용은 정보교환의 방법을 효율적으로 발전 시켰다. 그러나 협업을 위한 시스템에서 공통된 주제를 포함한 여러 문서가 생성되어 정보 공유의 힘을 분산 시키는 단점이 대두 되었다.

이러한 단점을 극복하기 위하여 본 논문은 협업문서 생성제어 시스템을 제안하였다. 파서(parser), 문서분류 시스템, 유사성측정 시스템으로 구성하였다. 협업문서에 대해 문서분류 시스템에서는 문서 분류시 문서의 변별력을 더욱 보장하고, 유사성측정 시스템에서는 이러한 변별력을 가지고 새로 만들어질 문서와 시스템내 저장된 문서의 유사성을 측정하여 문서를 순위화 하였다.

결과적으로 본 논문에서 제안한 시스템은 문서생성을 제어함으로써 시스템내의 정보 공유의 힘이 분산되는 것을 방지하여 각 분야의 전문가들의 원활한 정보공유와 지식창출을 효과적으로 할 수 있게 한다. 향후 시스템의 성능향상을 위해 문서분류 시스템에서 사용되는 더 좋은 알고리즘의 개발과 문서 분류의 기준이 되는 용어사전을 동적으로 구축에 대한 연구가 필요하다.

참고문헌

[1]http://no-smok.net/ns/moin.cgi/_bd_ac_bf_ee_c0_a7_c5_b0_bc_d2_b0_b3.

[2] http://no-smok.net/ns/moin.cgi/HelpContents.

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

[4] 김명철외 5명, 최신정보 검색론, 홍릉과학출판사.

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

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

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