지능형 E-mail 지식관리시스템 설계

박시일[0] 김두현 김용성
전북대학교 컴퓨터정보학과
{mysiil[0], kdh}@mail.chonbuk.ac.kr, yskim@moak.chonbuk.ac.kr

Designed of Intelligent E-mail Knowledge Management System

Si-Il Park[0] Doo-Hyun Kim Yong-Seong Kim
Dept. of Computer Information, Chonbuk University

요 약

본 논문에서는 E-mail을 적용한 지능형 E-mail 지식관리시스템을 제안하고자 한다. E-mail은 사용자에게 익숙하고, 정형화된 정보로 표현이 쉽고, 이미 많이 구축되어 있는 시스템이다.

이러한 E-mail의 정보를 활용하여 사용자에 따라 지식을 평가하고, 지식그룹 생성이 가능한 지식 관리 시스템을 설계한다. 이를 위해서 클러스터링을 이용해 지식간의 유사 정도에 따라 유사한 지식을 그룹화 시키는 지식그룹(Knowledge Group) 생성 알고리즘을 제안하고 사용자의 선호도(preference)를 반영하기 위해 사용자 프로파일(User Profile)을 설계하고, 사용자의 선호도에 적합한 지식을 검색하는고리즘을 제안한다.

1 . 소개

오늘날 우리 사회에서는 방대한 양의 문서가 지속적으로 생성된다. 이렇게 생성된 문서들은 다양한 정보를 저장하고 있기 때문에 이를 효과적으로 관리하기 위한 여러 가지 정보 시스템인 전자 우편(E-mail), 그룹웨어(Groupware), 문서 관리 시스템(DMS : Document Management System) 등이 제시되고 있다.[1][5]

각각의 정보 시스템에서 독자적인 체계에 의해 저장된 문서들을 지식 관리의 관점으로 생성된 계층구조에 의해 접근할 수 있도록 하였으며, 검색 기능을 통해 사용자가 원하는 문서들을 공유할 수 있도록 하고 있다.

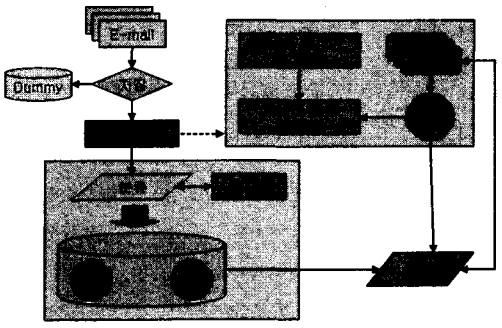
그러나 현재의 지식 관리 시스템이 제공하는 지식 공유에 대한 지원은 여러 가지 미흡한 점이 있다. 지식 공유를 효율적으로 하기 힘들고 시스템 관리자의 관점으로 지식을 관리하고 있다.

본 논문에서는 사용자 위주의 지식 관리 시스템을 위해, 사용자의 관심도를 나타내는 사용자 프로파일(User Profile)을 설계하고, 사용자 프로파일에 의한 자동 지식 검색 알고리즘과 유사한 지식을 그룹화 시키는 지식그룹을 생성하기 위한 알고리즘을 제안하고자 한다.

2 . 시스템 구조

현재 거의 모든 기업이나 공공기관에서 업무 및 정보에 대한 처리를 E-mail를 통하여 빠르게 전달하고 있다. 따라서 E-mail을 이용해서 지식관리시스템을 구축할 경우 추가적인 소프트웨어의 구축이 간단하고, 구조적 정보의 표현이 쉽고, 데이터 관리가 쉬워진다.

본 논문에서 설계한 지식관리 시스템은 [그림 2]와 같다.



[그림 2] E-mail을 이용한 지식 관리 시스템

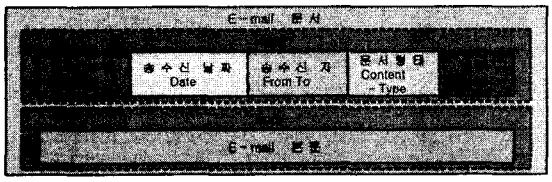
[그림 2]의 부분별 역할은 다음과 같다.

- 지식 추출기 : 사용자가 저장한 메일에서 각 데이터를 추출한다.
- 키워드 추출 : 메일의 제목과 내용에서 특정 용어의 발생 빈도가 높으면 높을수록 그 용어는 메일의 내용을 잘 표현하는 용어일 수 있다. 형태소 분석기를 이용하여 추출한 용어의 빈도를 이용해서 키워드를 추출한다.
- 사용자 프로파일 : 사용자의 관심 분야를 알기 위해 형태소 분석기에 의해 추출된 키워드를 통해 작성한다.
- 지식 분류 : 내용을 기반으로 유사한 메일을 그룹화해서 지식그룹을 생성한다. 지식그룹을 생성하는 이유는 관련성이 깊은 문서를 그룹화해서 검색 시 시스템 부하를 줄이고, 지식 활용을 효율적으로 하기 위해서이다.
- 자동 검색 : 저장된 문서에서 사용자의 프로파일이나 질의어에 의한 검색된 결과를 사용자에게 제공한다.

3 . E-mail의 정보 구조

기존의 지식관리 시스템과 E-mail을 이용한 지식관리 시스템의 차이점은 정보의 생성에 있다. 기존의 방법은 정보를 이용자 자신이 직접 생성하지만, E-mail을 이용한 지식 관리 시스템은 다른 사람이 생성한 정보를 활용한다는 것이다. 그렇게 때문에 정보를 관리하기 위해 필요한 생성일, 생성자, 정보 형태 등의 정보를 E-mail의 메시지에서 추출해야 한다.

[그림 2]는 E-mail의 원본메시지 구조를 head 부분과 body 부분으로 구분한 것으로, 원본메시지 구조를 살펴보면 지식관리시스템을 구축하기 위한 충분한 정보를 가지고 있음을 알 수 있다.



[그림 2] E-mail 문서 구조

원본메시지의 정보는 <표 1>과 같이 Data, Content-Type, Received, From, To와 같은 정형화 정보와 Subject, Content와 같은 비정형화 정보로 구분할 수 있다.

정형화 정보			비정형화 정보	
날짜	형태	문자열	문자열	
Date	Content-Type	Received From To	Subject	Content

<표 1> E-mail 정보 형태

정형화 정보는 정보의 표현이 명확한 것으로 관리나 분류가 단순한 질의에 의해서 가능하다. 그러나 비정형화 정보는 그 정보를 파악하기 위해서 먼저 정형화 정보로의 변환이 필요하다.

4. 지식그룹과 지식 평가

E-mail을 내용에 따라 분류하기 위해서 본 논문에서는 클러스터링 기법을 사용해서 메일간의 유사도를 측정한다. 클러스터링 기법은 E-mail에 부여된 키워드나 또는 기계적으로 추출된 키워드를 E-mail 내용의 식별요소로 삼아 E-mail 간의 유사도가 기준치 이상일 때 클러스터가 형성된다.[2][3]

4.1 지식그룹 생성

형태소 분석기에 의해서 추출된 키워드는 E-mail의 내용을 대표할 수 있는 중요한 요소로 임의의 E-mail 간에 공통되는 키워드가 많이 존재할수록 E-mail 간의 유사도가 높다고 볼 수 있다.

E-mail 간의 유사 정도를 측정하기 위해서 E-mail의 제목과 내용에서 추출한 키워드를 사용하고, 키워드의 발생 빈도를 가중치로 적용한다. 발생 빈도를 가중치로 적용할 때 E-mail마다 내용의 크기가 다르기 때문에 E-mail의 내용이 많을수록 특정 키워드가 자주 발생한다고 추정된다. 따라서 키워드의 발생 빈도를 각 E-mail의 내용 양에 따라 다른 가중치를 적용한 유사 측정공식은 다음과 같다.

두 E-mail X와 Y에서 추출한 키워드의 집합

X = Tx1, Tx2, ..., Txn, Y = Ty1, Ty2, ..., Tyn,

키워드에 따른 발생 빈도 집합

X = Wx1, Wx2, ..., Wxn, Y = Wy1, Wy2, ..., Wyn,

X와 Y의 키워드의 합집합

X Y = T1, T2, ..., Tn 일 때,

$$\text{유사도 (Similarity)} = \frac{\sum_{n=1}^{n}(T_n \times W_{xn} \times D_x) \times (T_n \times W_{yn} \times D_y)}{\sqrt{\sum_{n=1}^{n}(T_{xn} \times W_{xn} \times D_x)^2} \times \sqrt{\sum_{n=1}^{n}(T_{yn} \times W_{yn} \times D_y)^2}} \quad \cdots(1)$$

여기에서, D : E-mail의 내용의 크기에 따른 가중치.

본 논문에서는 KMS를 개발하는 업체의 E-mail 서버에서 78개의 E-mail 중 임의의 E-mail 9개를 선택하였다. 그리고 [4]의 형태소 분석기를 이용하여 각 E-mail에서 추출한 키워드와 발생 빈도는 <표 2>와 같다.

E-mail	키워드 (발생 빈도)	줄수
E1	컴퓨터(4), 구매(4), 기종(3), 운영체제(2)	20
E2	구매(4), 프린터(3), 자료(3), 성능(2), 가격(2)	18
E3	판매(6), KMS(3), 설명회(3), 기술(2)	20
E4	워크샵(3), DB(2), 공유(2), 향상(2), 기종(2)	16
E5	검색(3), 판매(3), 유지보수(2), 계약(2)	14
E6	구매(6), 필터(5), 순위(5), 검색(3)	32
E7	오라클(4), 구매(4), 컴퓨터(4), 가격(3), 서비스(2)	28
E8	지식(5), 워크샵(5), 분류(4), 공유(2), 정보(2)	24
E9	연봉(5), 평가(4), 능력(4), 인상(2), 검색(2)	24

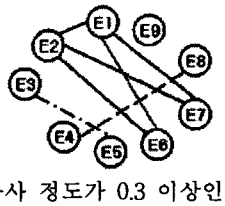
<표 2> 제목과 내용에서 추출한 정보

추출된 키워드와 발생 빈도 그리고 E-mail의 전체 줄 수를 이용하여 E-mail간의 유사도를 (1)의 공식을 적용하여 측정한 결과는 표 3과 같다.

	E1	E2	E3	E4	E5	E6	E7	E8	E9
E1	1	0.37	0	0.19	0	0.34	0.60	0	0
E2	0.37	1	0	0	0	0.39	0.43	0	0
E3	0	0	1	0	0.46	0	0	0	0
E4	0.19	0	0	1	0	0	0	0.44	0
E5	0	0	0.46	0	1	0	0	0	0.15
E6	0.34	0.39	0	0	0	1	0.23	0	0.08
E7	0.60	0.43	0	0	0	0.23	1	0	0
E8	0	0	0	0.44	0	0	0	1	0
E9	0	0	0	0	0.15	0.08	0	0	1

<표 3> 가중치를 고려한 E-mail간의 유사 정도

본 실험에서는 E-mail들의 유사 정도가 0.3 이상인 키워드들로 구성된 지식그룹은 그림은 [그림 3]과 같다.



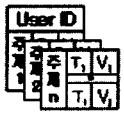
[그림 3] 유사 정도가 0.3 이상인 지식그룹

KL1 = E1, E2, E6, E7, K1 = 구매*, 컴퓨터, 가격

KL2 = E3, E5, K2 = 판매*
KL3 = E4, E8, K3 = 워트삽*, 공유
E9는 어떠한 E-mail과도 유사성이 없다.

4.2 사용자 프로파일

사용자 프로파일(User Profile)을 이용하면 많은 정보 중에서 사용자의 관심에 적절한 정보의 추출이 가능하다. 따라서 사용자는 새롭게 공유된 지식 중 자신과 관련이 있는 지식을 자동으로 추출할 수 있다.
사용자 프로파일의 구성은 [그림 3]과 같이 사용자 ID와 사용자의 관심분야의 용어(Term)들과 0에서 1사이의 용어에 대한 선호도 값(Value)으로 구성된다. T는 E-mail의 제목과 내용에서 추출한 키워드들로 구성된다.



User ID : 사용자 프로파일 구별
주제 : 사용자의 관심분야
T : 사용자 관심 용어
V : 어휘에 대한 선호도 값

[그림 4] 사용자 프로파일의 구조

사용자 프로파일의 갱신은 받은 E-mail에서 사용자가 문서를 저장하거나 질의어에 의한 정보 검색 시 발생하며, 사용자 프로파일 갱신은 두 가지 경우가 존재한다.
① 키워드가 사용자 관심 용어에 존재하는 경우
Vnij = Vnij +Vnj/n × C ??? (2)
② 키워드가 사용자 관심 용어에 존재하지 않는 경우
Vnij = C×0.01(기본값)??? (3)
단, n : T의 개수, C : 반복 횟수
Vnij : n번째 주제의 i 번째 용어에 대한 선호도 값
Vnj : 주제k에 존재하는 선호도 값
선호도 값을 0에서 1 사이로 나타내기 위해서 특정 관심 용어에 대한 선호도 값이 1보다 커질 경우, 그 값이 1로 되는 가중치를 전체 선호도 값에 곱한다.

4.3. 사용자 선호도를 반영한 지식 평가

한 집단의 지식을 공유 관리하는 지식 관리 시스템에는 방대한 양의 지식이 축적된다. 사용자가 많은 지식을 효율적으로 이용하기 위해서 새롭게 생성된 지식을 각 사용자의 관심도에 따라 평가해서 관련성이 높은 지식을 자동으로 검색해서 사용자에게 보여주어야 한다.
본 논문에서 제안하는 사용자별 지식 평가 알고리즘은 다음과 같다.

1. 새롭게 생성된 지식 추출
입력 : 사용자의 마지막 접속 날짜
출력 : 새롭게 생성된 지식

```
Function Select_Knowledge()
  DB_Connect()  //지식 데이터베이스에 연결
  //사용자의 마지막 접속 날짜 이후의 지식 추출
  Select_Knowledge(Last_Connect_Date)
```

2. 키워드와 발생 빈도 추출
입력 : 새롭게 생성된 지식
출력 : 각 E-mail의 키워드와 발생 빈도

```
Function Keyword_Count()
  for(i=0; i <= n; i++)
    for(j=0;  ; j++)
      //E-maili에서 키워드 Tj 추출
      Extract_Keyword(Tij)
      //키워드 Tj에 대한 발생 빈도 계산
      Count(Cij)
```

3. E-mail과 사용자와 관련성 계산
입력 : 사용자 프로파일, E-mail의 키워드, 발생빈도
출력 : 사용자에게 관련성이 높은 E-mail

```
Function Similarity_Degree()
  for(i=0; ; i++)
    //유사도 계산
    Degree = S( Tij, Cij, Tij, Vj)
    //유사도가 일정치 이상일 때 사용자에게 지식 제공
    If (Degree >= α)
      Show_User()
    Degree = 0
```

6 . 결론

본 논문에서는 효율적인 지식 관리 시스템을 위해, 이미 사용자에게 익숙한 E-mail을 사용하였다. 사용자 위주의 지식 관리 시스템을 위해 사용자의 선호도를 나타내기 프로파일을 설계하였고, E-mail에서 추출한 키워드를 이용해서 프로파일을 갱신한다.
또한, 사용자 프로파일에 의한 지식 평가 알고리즘과 유사한 지식을 그룹화 시키는 지식그룹(Knowledge Group)의 생성 알고리즘을 제안하였다.
본 논문에서 제안한 지식관리 시스템을 사용하면 이미 구축되어 있는 E-mail 시스템을 사용하기 때문에 별도의 하드웨어나 소프트웨어가 거의 필요 없고, 문서를 공유? 관리할 수 있는 구조화를 쉽게 할 수 있다.
향후 연구로는 E-mail 정보를 XML 구조의 문서로 자동 변환하여 저장하는 시스템과 본 논문에 제안한 알고리즘을 적용한 시스템을 구현하고자 한다.

7 . 참고 문헌

[1]. Alistair MacFarlane, Heriot-Watt University, "Information, Knowledge and Learning," Vol. 52, No. 1, pp. 77-92, 1998.
[2]. Mary Sumner, "Knowledge Management: Theory and Practice", Association for Computing Machinery, 1999.
[3]. 정영미, 정보검색론, 구미무역(주) 출판사, 1993
[4]. 강승식, 이하규, "한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사 기능", 한글 및 한국어 정보처리학회 학술발표논문집, 1998.
[5]. 이상진, "효과적 지식 공유를 위한 동적 폴더의 구현", 서울대학교 석사 학위 논문, 2000, 2