2002년 한국정보처리학회 춘계학술발표논문집 제9권 제1호

요약 문서 기반 문서 클러스터링

오형진, 고지현*, 안동언, 정성종
전북대학교 컴퓨터, 정보통신*공학과

e-mail : hyungjin@duan.chonbuk.ac.kr

Document clustering based on summarized document using K-means algorithm

Hyung-Jin Oh, Ji-Hyun Ko*, Dong-Un An, Sung-Jong Chung
Dept. of Computer, Information-Communication* Engineering, Chonbuk National University

요 약

정보검색 시스템에서 문서 클러스터링 기법은 사용자 질의에 대하여 검색된 문서를 문서간의 관련도에 따라 클러스터로 구성하고 사용자에게 검색 결과로 보여주는 것이다. 본 논문에서는 사용자의 질의에 대하여 검색된 문서를 자동 문서 요약기를 통해 얻은 요약 문서와 문서 전문을 문서들간의 유사도를 기반으로 동적으로 클러스터링 한다. 구현한 시스템의 클러스터링 효과를 검증한 결과 검색된 문서 전문을 클러스터링 한 방식에 비해 요약 문서를 클러스터링 한 방식이 정확률 측면에서 더 나은 성능을 보였다.

1. 서론

인터넷의 발달과 이에 따른 정보의 폭발적 증가로 온라인 텍스트나 전자화된 텍스트 문서의 양이 크게 증대되고 있다. 그러나 이러한 정보의 홍수로 인하여 사용자들로 하여금 필요한 정보를 쉽게 얻을 수 있는 것은 아니다. 현재의 정보검색 서비스들은 주로 대량의 웹 문서들 중에서 사용자 요구에 가장 적합한 문서들을 가능한 빠른 시간 내에 찾아 주는 것을 목적으로 검색 서비스를 제공하고 있다.

사용자에게 주어지는 정보검색 결과는 대부분의 경우 단순히 단어 빈도에 기초한 적합도에 따라 나열식으로 제시된다. 이 경우 사용자가 생각하는 적합도와는 다른 순서로 그 결과가 주어지는 경우도 있고 전혀 다른 의미를 갖는 문서들이 결과집합에 포함되는 경우도 있다. 따라서, 사용자들은 자신이 진정으로 원하는 문서를 찾기 위해, 이러한 검색결과 리스트를 순차적으로 살펴보아야 한다는 문제를 가지고 있다. 그러므로 정보검색에서는 사용자에게 질의에 적합하다고 판단되는 문서들을 빠르게 찾아주는 방법외에 사용자들이 검색 결과로 나온 문서 집합들의 의미를 빠르게 파악할 수 있는 방법에 대한 연구가 이루어져야 한다. 왜냐하면 다양한 주제에 관련된 텍스트를 모두 검색하고 조직화하는 것은 많은 시간과 노력을 필요로 하는 작업이다. 컴퓨터를 이용한 문서 자동분석에 대한 요구가 증대되고 있는데 문서 클러스터링은 그러한 요구를 충족시키는 방법중의 하나이다.

2. 관련 연구

문서 클러스터링이란, 특정 문서집합 내에 있는 각 문서들간의 유사도를 측정하여 유사한 문서들을 그룹화하는 것을 의미한다. 문서들간의 유사도는 각 문서가 갖는 특징을 비교하여 계산하게 되는데 일반적으로 문서에 포함되어 있는 단어의 빈도수를 그 특징으로 사용한다. 문서 클러스터링 기법은 클러스터를 구성해나가는 방법에 따라 계층적 클러스터링과 비계층적 클러스터링으로 나누어 볼 수 있다.

계층적 클러스터링은 비계층적인 방법에 비해 클러스터링 시간이 느리지만 보다 정확한 클러스터링이

수행된다는 장점을 갖는다. 비계층적 클러스터링은 임의로 선택된 초기 클러스터로부터 문서를 클러스터에 재배치하는 작업을 반복적으로 수행하여 최종 클러스터를 형성하는 방법으로 계층적 클러스터링에 비해 클러스터링 시간은 빠르지만 검색 효율이 떨어지고 문서의 입력 순서에 따라 클러스터링 결과가 달라진다는 단점을 갖는다. [6]

본 논문의 범위는 정보검색 시스템에서 사용자의 질의에 대한 검색 결과인 문서를 자동 문서 요약기를 이용하여 요약한 후, 요약된 형태의 문서를 비계층적 방법으로 클러스터링한 결과와 검색 문서의 전문을 클러스터링한 결과 성능을 분석하는 것이다. 실험 결과 검색된 문서 전문을 클러스터링 한 방식에 비해 요약 문서를 클러스터링 한 방식이 정확률 측면에서 더 나은 성능을 보였다.

본 논문의 구성은 다음과 같다. 3 장에서는 요약 문서 기반 K-means 클러스터링을 수행하는 시스템의 구성 및 모듈에 대하여 기술을 하며, 4 장에서는 실제 처리 예와 실험 결과를 보여주며, 끝으로 5 장에서는 결론을 맺겠다.

3. 시스템 구성 및 모듈

본 논문에서 구현한 시스템은 크게 세가지 모듈로 구성되어 있다. 먼저 문서 정보 모듈은 정보 검색 엔진으로부터 검색된 문서를 자동 문서 요약기를 거쳐 생성한 문서로 각각의 요약 문서는 DID, TID, TF, DF 로 구성되어 있다. DID 는 문서 번호, TID 는 검색된 문서에서 중요 단어의 번호, TF 는 현재 문서(DID)내에서의 단어 빈도수이며 DF 는 전체 검색된 문서중 TID 가 출현한 문서 개수이다. 두 번째 모듈은 클러스터링을 담당하는 부분이다.

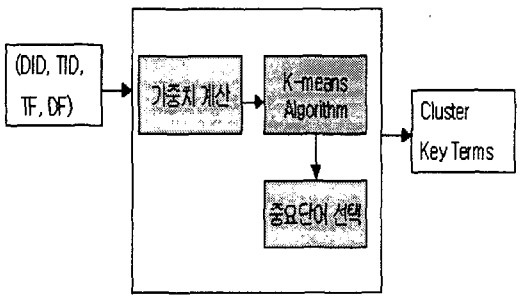

그림 1. 시스템 구조

단어의 가중치 계산 방법은 식 (1)과 같다.

$$v_{ij} = Tf_{ij} \cdot idf(w_{ij}) \qquad (1)$$

여기에서,

v_{ij} : j 번째 문서의 i 번째 단어의 가중치

$Tf_{ij} = \frac{tf_{ij}}{tf_{ij}+2}$: j 번째 문서의 i 번째 단어의 단어 빈도수

$idf(w_{ij}) = \max(M, \log\frac{N}{df_{ij}})$

j 번째 문서의 i 번째 단어의 역문서 빈도수

df_{ij} : j 번째 문서의 i 번째 단어의 문서 빈도수

N : 검색된 전체 문서 빈도수

구현한 K-means 알고리즘은 다음과 같다.

1.클러스터 수 K 개를 선택한다

2.Proto-Centroid 를 선택한다.

3.문서-클러스터간 거리를 계산한다.

$$\arg\min_{\substack{i=1,n \\ j=1,k}} dist(\vec{d_i}, \vec{c_j})$$

4.문서 d_i 를 클러스터 c_j 에 할당한다.

$$d_i \in \bigcap_{i=1,n} G_{c_j} \; if\; dist(d_i,c_j) < dist(d_i,c_l)\; for\; all\; l = 1,2,\cdots,k \quad l \neq j$$

5.클러스터 중심을 재계산한다

$$\vec{c_j} = \frac{1}{|c_j|}\sum_{l=1}^{|c_j|}\vec{d_l}$$

6. 새로 생성된 클러스터 중심과 이전에 생성된 클러스터 중심과의 거리가 임의의 값 이상이면 3 으로 가서 반복한다.

클러스터와 문서의 거리 계산에 Euclidean distance 를 사용하였다.

$$(\vec{d_i}, \vec{c_j}) \; : \; \sqrt{\sum_{k=1}(d_{ki} - c_{kj})^2}$$

세 번째 모듈은 실제로 검색된 문서들의 요약 문서를 K 개의 클러스터로 그룹화 한 후 각 클러스터의 대표 단어를 추출하는 모듈이다.

4. 실험 및 결과

본 논문에서 실험은 102 개의 TREC 문서[8]를 대상으로 하였다. K-means 알고리즘에서 생성할 클러스터의 수는 10(k=10)으로 선택하였다. 구현한 시스템에 대한 검색된 문서의 요약 문서는 다음과 같다.

<원문>

```
<DOC>
<DOCNO>WTX001-B02-1</DOCNO>
<DOCOLDNO>IA001-000000-B026-182
</DOCOLDNO>
…
<title>
Cannot set an installable partition with FDISK
```

```
</title>
…
<h2>Symptoms</h2>
When Advanced Installation is chosen for the installation of Warp, FDISK
thinks that <code>"your partition mapping may be corrupt"</code>. Partitions
can only be deleted but this won't help. As a result, you can only
use Easy Installation and choose the default partition. This may not be what
you want.

…
Partitions on your harddisk (for example Linux' ext2fs) may be misaligned so that
it confuses FDISK. This is probably a bug in Warp.
…
The FDISK included with OS/2 2.1 (or Warp Beta II) does not have this problem. If you
can't get this FDISK, contact me.
<p>
There are at least four possible solutions to this problem.
<p>
<OL>
<LI>If you have the floppy version of Warp, backup the disk containing FDISK
(<b>Diskette 1</b>). Copy the FDISK of 2.1 onto that disk, replacing Warp's
FDISK.<p>

…
```

각 요약 문서는 termeid, term, tf, df로 구성되어 있다.

<요약 문서>

```
<DOC>
<DOCNO>WTX001-B02-1</DOCNO>
43      problem         3       23616
3       partit  6       506
44      harddisk        4       28
45      linux   1       339
46      ext2f   1       2
47      misalign        1       42
48      confus  1       3235
4       fdisk   15      38
58      floppi  4       842
63      copi    7       32235
126     x:Wos2imageWdisk_1Wsysinst2.ex  1  1
127     binari  1       871
93      file    3       18268
128     editor  1       7441
129     debug.ex        1       3
125     sysinst2.ex     3       2
107     fdisk.com       3       4
62      diskett         5       745
131     fdork.
```

구현한 시스템에 대한 문서의 전문은 다음과 같다. 각 문서의 전문은 termeid, term, tf, df로 구성되어 있다.

<문서 전문>

```
<DOC>
<DOCNO>WTX001-B02-1</DOCNO>
<DOCLENGTH>294</DOCLENGTH>
1       set     4       25786
2       instal  19      8411
3       partit  6       506
4       fdisk   15      38
1       set     4       25786
2       instal  19      8411
3       partit  6       506
4       fdisk   15      38
5       world   1       32208
6       australia       1       4353
7       europ   2       4672
8       north   2       13072
9       america         2       11020
10      singapor        1       1224
7       europ   2       4672
11      austria         1       859
12      birmingham      1       838
13      budapest        1       272
14      franc   1       3476
15      itali   1       1804
16      london  1       5303
17      teamos/2        2       123
18      germani         1       3365
8       north   2       13072
9       america         2       11020
19      berkelei        1       1542
20      chattanooga     1       389
21      minnesota       1       2784
22      nova    1       781
23      scotia  1       544
24      ontario         1       1854
25      pennsylvania    1       2429
17      teamos/2        2       123
…
</DOC>
```

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
요약문	2	1	1	1	15	1	1	72	7	1
정확률	100	100	100	100	86.7	100	100	65	100	100

표 1.클러스터에 할당된 문서수와 정확률(요약문)

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
전문	2	1	5	2	84	2	1	1	2	2
정확률	0	100	60	0	49	0	100	100	0	0

표 2. 클러스터에 할당된 문서수와 정확률(전문)

실험 결과(표 1,표 2)에 의하면 요약 문서의 정확률이 전체 문서를 클러스터링한 결과보다 더 나은 성능을 보이는데 요약 문서의 경우가 문서 전문 보다 일관성 있는 데이터를 가지고 있기 때문이다. 또한 전문을 클러스터링 한 경우에는 정확률이 0%인 경우가 존재하는데 두 문서의 내용이 아주 다르기 때문이며 평가는 사람에 의하여 판정하였다.

5. 결론

본 논문에서는 정보검색 시스템에서 사용자 질의에 따라 검색된 문서를 자동 문서 요약기를 통해 얻은 요약 문서와 문서 전문을 문서들간의 유사도를 기반으로 동적으로 클러스터링 하였다. 구현한 시스템의 클러스터링 효과를 검증한 결과 검색된 문서 전문을 클러스터링 한 방식에 비해 요약 문서를 클러스터링 한 방식이 정확률 측면에서 더 나은 성능을 보였는데 요약 문서의 경우 문서 전문 보다 더욱 일관된 데이터를 가지고 있기 때문이다.

향후 연구로는 요약문의 크기를 변화시킴에 따라 클러스터링에 미치는 영향에 대하여 살펴보겠다.

참고문헌

[1] A. Leuski and J. Allan. "*Improving interactive retrieval by combining ranked lists and clustering.*" In Proceedings of RIAO'2000, pages 665--681, April 2000.

[2] Ray R. Larson' s Lecture Notes of Principles of Information Retrieval.

[3] Prabhakar Raghavan' s Lecture Notes of Principles of Information Retrieval.

[6] 김태현, 맹성현. " 계층구조를 이용한 문서 클러스터 제목의 자동생성" 2001 년도 제 13 회 한글 및 한국어 정보처리 학술대회, pages 163-170. 2001. 11

[7] 박진우, 고영중, 서정연, " 문서 요약 기법을 이용한 자동 문서 범주화" , 2001 년도 제 13 회 한글 및 한국어 정보처리 학술대회, pages 138-145. 2001. 11

[8] http://trec.nist.gov