MMR, 클러스터링, 완전연결기법을 이용한 요약방법 비교

유준현, 변동률, 박순철
전북대학교 정보통신학과
전화 : 063-270-2467 / 핸드폰 : 018-608-5946

Comparisons of MMR, Clustering and Perfect Link Graph Summarization Methods

Jun Hyun Lyu, Dong Ryul Byun, Soon Cheol Park
Dept. of Electronic & Information Engineering, Chonbuk University
E-mail : soso7@internet.chonbuk.ac.kr

Abstract

We present a web document summarizer, simpler more condense than the existing ones, of a search engine. This summarizer generates summaries with a statistic-based summarization method using Clustering or MMR technique to reduce redundancy in the results, and that generates summaries using Perfect Link Graph. We compare the results with the summaries generated by human subjects. For the comparison, we use FScore. Our experimental results verify the accuracy of the summarization methods.

I. 서론

인터넷의 사용자의 증가와 전자문서의 증가에 따라서 사용자들이 접근할 수 있는 문서의 양이 증가하고 있다. 이러한 문서 중 자신이 원하는 것을 찾기 위해서는 문서의 전체의 내용을 읽어야 하지만 그것은 쉽지 않다. 때문에 문서의 내용을 쉽게 파악하기 위해서는 중요한 문장의 일부를 보여주는 자동요약이 필요하다.

본 논문에서는 정보검색의 기술 중 문서요약에 대한 연구를 수행하였다. 문서 요약의 경우 대부분 통계기반을 이용하여 요약을 한다. 그러나 이러한 방법은 요약의 정확성이 떨어지고 전체의 내용을 대표하지 못하는 경우가 많다. 그 외 의미기반 요약 방법과 문맥을 이용한 요약 등이 연구 중에 있지만 구현이 쉽지 않고 요약에 대한 신뢰성이 떨어진다. 이를 개선하기 위한 통계기반의 방법으로, 중복되는 문장을 제거하는 MMR(Maximal Marginal Relevance)기법과 본 논문에서 제안하는 클러스터링을 이용한 요약, 그리고 문장간의 관계를 이용하는 관계거리요약과 비교하였다.

II. 요약방법

2.1 MMR을 이용한 요약

MMR 알고리즘은 정보검색시스템의 검색결과에서 중복내용을 제거하여 결과를 보여주는데 사용하고 있다. 본 논문에서는 이러한 MMR의 중복성 제거 특성을 이용하여 중복이 적고 많은 정보를 포함하는 것을 요약 문장으로 선택하였다.

문서내의 문장에서 사용되는 단어의 가중치의 계산 방법은 식 (1)과 같다. 단어의 가중치, s_{ij}는 j문장에서 i번째 있는 단어, w_{ij}의 가중치이다.[1]

$$s_{ij} = f_{ij} \cdot idf(w_{ij}) \cdot P(w_{ij}) \qquad (1)$$

여기서,

$$f_{ij}=\frac{freq_{ij}}{freq_{ij}+2},\ idf(w_{ij})=\log\frac{N}{n_{ij}},\ P(w_{ij})=\begin{pmatrix}2.0 & high\\ 1.5 & important\\ 1 & others\end{pmatrix}$$

이다.

식 (1)에서 $freg_{ij}$는 문서에서 단어 w_{ij}의 빈도수이며 $freg_{ij}$는 요약할 문서 내에 있는 단어 중 최대빈도수를 갖는 단어 w_{ij}의 빈도수이다. 따라서 f_{ij}는 $freg_{ij}$를 정규화한 단어 w_{ij}빈도의 정도를 나타내며 0부터 1사이의 실수값이다. 역문서빈도수 $idf(w_{ij})$의 계산식 중 N은 본 시스템에서 문서의 총 수이다. n_{ij}는 단어 w_{ij}가 출현한 문서의 수이다.

각 문서 내의 문장을 문장 벡터로 표현할 때 식 (2)와 같이 표현할 수 있다. 이 문장 벡터는 문장의 중요도를 계산할 때와 문장과 문장사이의 유사도를 계산할 때 사용되어진다.

$$\vec{S_i}=(s_{1,i}, s_{2,i}, \cdots, s_{n,i}) \quad (2)$$

문장의 중요도는 문장 벡터 내에 있는 단어들의 평균값으로 표현된다. 식 (3)은 문장의 중요도이다.

$$|\vec{S_j}|=\frac{\sum_{i=1}^{size\ of\ \vec{S_j}} s_{i,j}}{size\ of\ \vec{S_j}} \quad (3)$$

기본적인 MMR 통계요약 알고리즘은 식 (4)와 같다. MMR을 적용하지 않았을 때는 가중치가 높은 단어를 포함하는 문장을 중복하여 선택하게 된다. 이것을 보완한 것이 본 논문에서 사용하는 MMR[3] 기법이다.

$$arg\ \max_{S_i\in R-A}\{(|\vec{S_i}|-\lambda\cdot\max_{S_j\in A}\cdot(sim(\vec{S_i},\vec{S_j}))\} \quad (4)$$

여기서 A는 요약 문장으로 선정된 문장의 집합이다. R은 문장의 중요도에 의해 정렬된 리스트이다. 유사도 함수, $sim(S_i, S_j)$는 코사인 유사도를 사용하여 선택 가능한 문장과 이미 추출된 요약 문장 집합, A 에 포함된 문장과 비교하였다. 임의 두 문장, S_i와 S_j 간의 코사인 유사도 계산은 식 (5)와 같으며 유사도 값은 0과 1사이의 실수이다.

$$sim(\vec{S_i},\vec{S_j})=\frac{\vec{S_i}\cdot\vec{S_j}}{|\vec{S_i}|\times|\vec{S_j}|}=\frac{\sum_{k=1}^{n} s_{ki}\times s_{kj}}{\sqrt{\sum_{i=1}^{n} s_{ki}^2}\times\sqrt{\sum_{m=1}^{n} s_{mj}^2}} \quad (5)$$

2.2 클러스터링을 이용한 요약

K-Means 클러스터링 알고리즘은 정보검색 분야에 많이 사용된다. 이 클러스터링 방법을 사용하여 관련 문서끼리 분할하고 이중에 중요한 문장을 선택함으로써 중복성을 제거하게 된다.

1. Choose k.
2. Select k initial centroids, cj, where 1≤j≤k.
3. for i = 1 to no. of sentences
4. for all j=1, 2, ···, k.
5. Compute dist(si, cj).
6. endfor
7. Select j for the minimum dist(si, cj).
8. Assign si to the jth cluster.
9. endfor
10. Recompute the new centroids for each cluster.
11. Check if (old centroids ≒ new centroids) then return.
12. else goto 3.
13. Choose the nearest sentences of the centroids in the clusters.

그림 1. 클러스터링 기법을 이용한 요약 알고리즘

문장간 일치하는 용어의 수는 적다. 그렇기 때문에 문장을 이용하여 초기 센트로이드를 설정하는데는 어려움이 있다. 본 논문에서는 2~3개의 문장을 초기 센트로이드로 설정하여 이 문제를 보완하였다. 초기 센트로이드 계산 방법은 식 (6)과 같다.

$$C_j^{initial}=\frac{\sum_i |s_i|\cdot\overrightarrow{SV_{s_i}}}{\sum_i |s_i|} \quad (6)$$

여기서 $|s_i|$는 문장의 용어수이고 $\overrightarrow{SV_{s_j}}$는 s_i의 문장벡터이다.

2.3 완전연결기법을 이용한 요약

자동요약은 사람이 한 요약과 얼마나 유사한지에 따라 정확도가 결정된다. 일반적인 인위적인 요약은 전체 내용을 파악하여 관련성이 있는 문장중 내용을 대표하는 문장으로 요약한다. 이런 인위적 요약의 특성을 이용한 방법으로 Salton의 Text relationship map이 있다.[4]

Salton의 요약 방법은 문장간의 관련성이 적은 문장을 제거한 후 문장간 사슬 관계에 의해 요약을 하게 된다. 그렇기 때문에 미미한 값, 즉 관련성이 적은 문장의 거리값은 적용되지 않는다. 이를 보안한 문장간의 거리값을 모두 합쳐서 계산한 방법으로 '도합유사도'가 있다.[5]

위 두 방법은 유사성이 높은 문장을 요약문으로 선택하는 방법을 사용하였다. 본 논문에서는 문장간의 유사도가 높은 경우만이 아닌 전혀 유사하지 않는 문장을 요약문으로 선택하여 결과를 평가하였다. 식 (7), (8)은 본 논문에서 사용한 요약의 알고리즘이다.

$$\arg\min_k\left\{\sum_{j=1}^{n} sim(S_i,S_j)^2\right\} \quad (7)$$

$$\arg\max_k\left\{\sum_{j=1}^{n} sim(S_i,S_j)^2\right\} \quad (8)$$

문장간의 유사도 계산 방법으로 Cosine similarity, Inner product, Euclidean distance를 이용하였다. Cosine similarity, Inner product 유사비교 방법에는 식(7)을 적용하고 Euclidean distance 유사비교 방법에는 식(8)을 적용하였다.

Ⅲ. 인위적 요약

실험문서로는 20개의 Korea Times 기사를 이용하였다. 각 문서는 평균 11.5개의 문장으로 실험 문서의 형평성을 고려하여 문서 내의 문장수가 11~13개인 문서를 선택하였다.

인위적 요약에 참가한 사람은 전북대학교 인문대 대학원생 5명과 본 요약시스템 개발자를 포함한 전북대학교 공과대 대학원생 4명이다. 각 실험대상자는 기사를 먼저 읽고 각각의 기사에 대해 가장 중요한 문장을 1개 선택하고 그 다음 문장의 중요성에 따라 차기 문장을 선택하도록 했다. 이렇게 해서 선택된 5개의 문장을 중요도에 따라 1에서 5까지 각각 점수를 부여했다. 그림 2는 인위적 요약방법에 따라 선택된 문장들의 평균 점수변화를 보인다. 그림 2에서는 대부분 뉴스기사의 중요한 문장이 앞에 위치한다는 것을 보여준다.

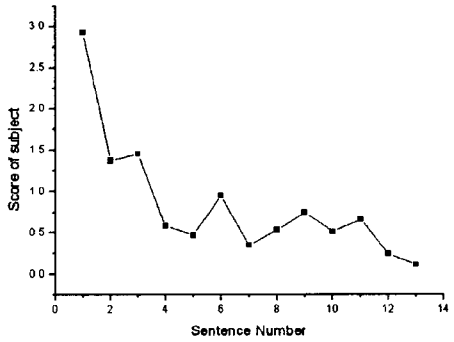

그림 2. 인위적 요약방법에 의해 선택된 문장들의 점수변화

Ⅳ. 실험결과

본 실험은 인위적 문서요약을 기준으로 하여 MMR, 클러스터링, 완전연결기법을 이용한 요약을 비교하였다. 비교 값으로는 재현율과 정확율의 정보를 포함하고 있는 FScore를 사용하였다

$$recall = \frac{|B|}{|A+B|} \qquad precision = \frac{|B|}{|B+C|},$$
$$FScore = \frac{2*recall*precision}{recall+precision} = \frac{|CS \cap HS|}{No.\ of\ summaried\ sentences} \tag{9}$$

여기서 CS는 자동요약으로 생성한 문장집합이고 HS는 인위적 요약의 문장 집합이다.

본 실험에서는 비교를 용이하게 하기 위하여 인위적 요약문서의 문장수와 통계 요약문서의 문장수를 같게 하였다. 따라서 재현율의 값은 정확율과 같게되며, 아울러 FScore의 값도 재현율이나 정확율의 값과 같게 된다.

그림 3은 MMR을 이용한 요약을 하였을 때 임계값(λ)에 따른 요약의 성능을 나타낸다. 그림에서는 λ =0.6일 때 좋은 결과를 보여준다.

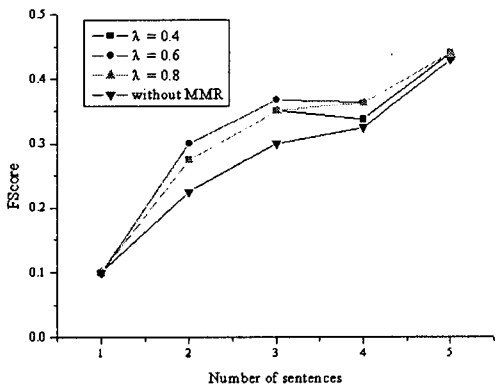

그림 3. 임계값(λ)과 요약문장 수에 따른 FScore

그림 4는 요약 문장수와 요약알고리즘, 완전연결기법의 유사도 비교방법에 따른 FScore값을 보여준다. 3문장 이상의 요약에서는 Without MMR, 즉 통계정보만을 이용하여 요약한 경우보다 다른 요약방법의 성능이 모두 우수했다. 특히 Inner Product와 Euclidean Distance 유사도 비교로 완전연결기법을 사용한 경우 1~2문장 요약에서 다른 요약방법보다 FScore값이 매우 높게 나왔다.

검색엔진에서 요약문은 보통 2~3개의 문장으로 이루어져 있다. 원문의 길이에 비하면 매우 적은 내용이다. 일반적인 통계기반 방법은 원문의 내용보다 적은 요약일 경우 그림2(Without MMR)와 같이 성능이 저하되는 것을 볼 수 있다. 완전연결기법을 이용하여 요약을 할 경우 적은 문장에서도 요약의 성능이 다른 방법보다 매우 우수하다. 때문에 웹문서와 같은 적은 수의 요약에서 완전연결기법 요약처럼 문맥을 이용한 요약방법을 사용하면 보다 정확한 요약문을 생성할 수 있을 것이다.

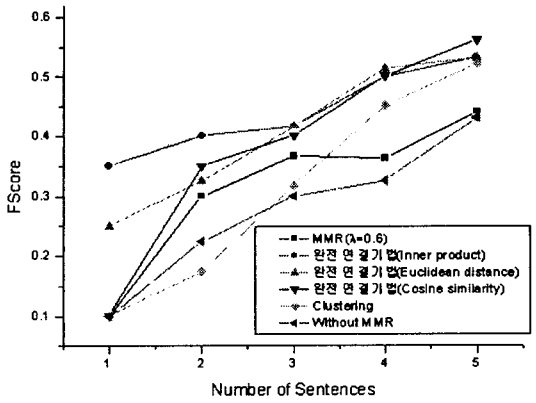

그림 4. 요약 문장수와 요약방법에 따른 FScore

Salton의 요약방법에서는 문장간의 유사도가 높은 것만을 이용하여 요약에 사용하였지만 본 논문에서는 문장간의 유사도가 낮은 값들도 모두 합하여 적용하였다.

기존의 도합유사도 방법은 모든 수치의 합을 정규화하지 않았다. 본 논문에서는 이 값을 제곱함으로써 정규화시켰다. 또한 Euclidean distance를 사용한 경우는 거리가 가장 먼 값을 요약문으로 채택함으로써 기존의 유사도가 높은 개념이 아닌 다른 개념으로 요약문 생성에 접근하였다.

식 (10)~(12)는 유사도 계산식을 나타낸다. 식 (10)은 Cosine similarity, 식 (11)은 Euclidean distance, 식 (12)는 Inner product를 나타낸다. Cosine similarity는 문장 내 포함하는 용어의 수가 많을 때, 즉 문장이 많은 정보를 포함하는 경우 분모값이 커지기 때문에 전체적인 유사도 값이 작게 나온다. 이에 반해 Euclidean distance의 경우 문장 내 포함하는 용어수가 많아지면 유사도 값이 크게 나오게 된다. 그렇기 때문에 본 논문에서는 이점을 착안하여 Euclidean에서 유사도 값이 클 때를 요약문으로 채택하였다.

$$sim(\vec{S_i}, \vec{S_j}) = \frac{\vec{S_i} \cdot \vec{S_j}}{|\vec{S_i}| \times |\vec{S_j}|} = \frac{\sum_{k=1} s_{ki} \times s_{kj}}{\sqrt{\sum_{i=1} s_{ki}^2} \times \sqrt{\sum_{m=1} s_{mj}^2}} \quad (10)$$

$$sim(\vec{S_i}, \vec{S_j}) = \sqrt{(\vec{S_i} - \vec{S_j})^2} = \sum_{k=1} \sqrt{(s_{ki} - s_{kj})^2} \quad (11)$$

$$sim(\vec{S_i}, \vec{S_j}) = \vec{S_i} \cdot \vec{S_j} = \sum_{k=1} s_{ki} \times s_{kj} \quad (12)$$

V. 결론

본 논문에서는 MMR, 클러스터링, 완전연결기법을 이용하여 요약을 수행하였다. 문장간의 의미, 즉 문맥을 적용한 완전연결기법이 가장 결과가 좋았고 클러스터링요약의 경우 문장수를 많이 뽑았을 때 결과가 좋았다. 클러스터링요약의 경우 요약 문장수가 적을 때 결과가 안 좋은 이유는 클러스터링한 후 대표문장을 선택할 때 단순히 통계적인 가중치가 높은 것을 뽑았기 때문이다. 선택하는 방법을 본 논문에서 제안한 완전연결기법을 이용하여 선택하게 되면 더 좋은 결과를 얻을 수 있을 것이다.

Euclidean을 이용한 완전연결기법은 거리의 값이 먼 경우, 즉 각 문서간의 연관성이 가장 없고 용어의 정보를 많이 포함한 내용을 요약문으로 선택한다. 이런 특성은 초기 센트로이드 설정에 유용하다. 센트로이드는 용어수가 많을수록, 거리가 떨어진 것을 채택할수록 클러스터링이 잘 생성된다.

정보검색에서 원하는 문서를 선택할 때 가장 최 상단에 원하는 문서가 위치한 경우는 적다. 이는 ranking을 정하는 방법에 있어 일반적인 통계적 방법을 사용하기 때문이다. 여기에 완전연결기법을 이용하여 검색된 문서 중 대표하는 문서 하나를 선택하게 되면 좀 더 정확한 ranking을 보여줄 것이다.

문장의 유사도 비교 방법 중 Inner product와 Cosine similarity에는 문제점이 있다. 한 문장 자체가 가지고 있는 정보가 적기 때문에 전혀 유사하지 않은 문장에서는 요약이 어렵다. 때문에 사전이나 유사어, 관련어를 이용한 용어확장 방법이 필요하다. 이런 방법을 이용하면 좀 더 정확한 요약문을 얻을 수 있을 것이다.

참고문헌

[1] 강상배, 한국어 문서의 통계적 정보를 이용한 문서요약 시시스템 구현, 부산대학교, 전자계산학과, 석사 학위 논문, 1998. 2.

[2] http://nlp.kookmin.ac.kr/ 국민대학교 강승식교수, 한국어 분석 모듈(HAM)

[3] Jaime Carbonell and Jade Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.

[4] G.Salton, A.Singhal, M.Mitra, and C.Buckley, Automatic Text Structuring and Summarization

[5] 김준홍, 도합유사도를 이용한 한국어 추출요약 시스템, 한국해양대학교, 컴퓨터공학과, 석사 학위 논문, 2000. 8.