제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

MMR을 이용한 질의기반 자동 문서요약

김금영*, 안동언*, 정성종*
전북대학교 컴퓨터공학과
e-mail:com94@duan.chonbuk.ac.kr

Query-Based Automatic Text Summarization Using MMR

Gum-Young Kim* Dong-Un An* Sung-Jong Chung*
Dept of Computer Engineering, Chonbuk University

요 약

정보검색 엔진들은 주어진 질의에 대해 방대한 양의 문서들을 검색해주며, 이 문서들은 질의와의 관련성에 따라 랭킹(Ranking)된다. 검색된 문서들중에 어떤것들은 그 내용이 서로 유사하여 사용자에게 필요 이상의 정보를 제공한다. 이는 질의와의 관련성(Relevance)만을 적용하고, 검색된 정보들간의 차별성을 고려하지 않은데서 비롯된다. MMR(Maximal Marginal Relevance)은 유사한 문서를 검색결과에서 배제할 수 있게 해주는 기법이다. MMR을 자동요약에 적용하면, 유사한 문장을 배제하여 상이한 정보들을 전달하는 질 높은 요약문을 생성할 수 있다.

본 논문에서는 MMR을 이용한 질의기반 자동 문서요약 시스템을 구현한다. 또한, MMR과 가중치 수식에 다양한 수치를 적용하고, 최적의 결과를 산출하는 수식을 제안한다.

1. 서론

최근 10여년간 인터넷의 이용은 가히 폭발적으로 늘어났으며, 이젠 정보검색이란 말은 평범한 말이 되어가고 있다. 또한, 사용자들이 정보검색을 일상처럼 사용하면서 정보검색엔진을 이용하는 수준이 높아져, 더 높은 질의 검색결과를 원하게 되었다.

정보검색 엔진들은 주어진 질의에 대해 방대한 양의 문서들을 검색해주며, 이 문서들은 질의와의 관련성에 따라 랭킹(Ranking)된다. 따라서, 같은 정보를 가진 문서들이 서로 비슷한 위치에 놓여있게 되며, 사용자들이 다른 정보를 가진 문서를 검색하기 위해서는 페이지를 이동하여야만 한다.

이는 문서들의 랭킹(Ranking)에 질의와의 관련성만을 부여하고, 랭킹된 문서들끼리의 차별성는 고려치 않은 까닭이다. 이렇게 검색된 문서들에 차별성을 부여한다면, 좀더 좋은 질의 검색결과를 얻을 수 있을 것이다. 즉, 사
용자들은 페이지를 이동하지 않아도 상이한 정보들을 가진 문서들을 검색할 수 있게 된다. 자동요약에 이러한 차별성을 적용한다면, 생성될 요약문에서 유사한 문장을 제거하여 좀더 많은 정보를 가진 요약문을 생성할 수 있다.

본 논문에서는 Carbonell[6]이 제안한 유사문장을 배제하는 기법인 MMR(Maximal Maginal Relevance)을 이용한 문서 자동요약 시스템을 구현한다. 요약방식은 문장 추출에 기반한다. 높은 질의 결과를 산출하기 위해서, MMR을 사용하지 않은 자동요약 시스템에서 문장을 추출하기 위한 가중치 수식에 다양한 수치를 적용한 후 실험하여 최적의 결과를 산출한 후 MMR을 적용한 시스템의 결과와 비교하였다.

실험에 사용된 문서집합은 논문100편을 사용하였으며, 시스템에서 생성된 요약결과는 논문의 저자가 작성한 요약과 유사도 비교를 통하여 평가한다.

본 논문은 2장에서 관련연구를 논하고, 3장에서는 본 논문의 문서 자동요약 시스템의 구조에 대해서, 4장은 MMR을 적용한 시스템의 실험결과에 대해서 5장에서는 결론을 내린다.

2. 관련연구

1960년대부터 연구되기 시작한 문서요약은 원본에서 중요한 내용을 선택과 생성에 의한 압축을 통하여 요약문을 생성하여 원본을 축소시키는 과정이라 할수 있다. [7]

문서 요약을 접근 방법으로 나누어 볼 때 문장추출 방식과 정보 추출 방식으로 나눌 수 있다. [5]

문장추출 방식은 요약 문장을 결정하기 위해 통계적인 방법, 위치나 단서단어, 문장 특성, 수사구조를 이

용하여 중요문장을 선택하여 요약문을 생성한다. [1]

정보 추출기반 문서 요약 시스템은 문서의 종류에 따라 추출되어야 할 개념들이 정해져 있고, 이러한 개념을 문서내에서 추출하여 주어진 템플릿을 채우는 방식으로 요약문을 생성한다.

또 다른 연구로 주어진 질의에 대해 의사 적합성 피드백, 제목, 문서의 첫 문장등을 이용하여 요약문을 제시하는 질의기반 문서요약 방법이 있다. [2]

본 논문은 고빈도 단어의 영향을 줄이기 위해 추가적인 상수를 사용하며 문장단위는 형태소 해석기가 추출하는 색인어수를 일정하게 포함하는 Passage로 한다. [5]

3. 시스템의 구조

질의기반 문서요약 시스템은 선처리부와 실행부로 나누어진다. 선처리부는 주어진 문서집합을 입력으로 받아 형태소 해석기(MORAN2001) [3]가 각각의 문서에서 색인어를 추출하는 모듈과 색인된 문서집합에서 통계치를 추출하는 모듈로 나누어진다. [그림2]

실행부는 주어진 문서에서 문장단위(Passage)로 가중치를 부여하는 모듈, 요약문 생성 모듈로 나누어 진다. 실행부에서는 주어진 질의에 대하여 선처리부에서 계산되어진 통계값을 각 문장단위에 가중치로 적용하여 요약문을 생성한다.

가중치는 단어 빈도수, 역문서빈도수를 부여한 후 MMR을 적용하여 계산한다. 요약문을 추출하는데에 있어서 단어 빈도수가 요약문의 질을 크게 좌우하지는 않는다. [1] Okapi TF는 단어 빈도수의 가중치에 끼치는 영향을 1.0이하의 수치로 제한할 수 있다.

$$TF_{ij} = \frac{tf_{ij}}{tf_{ij} + 2.0} \qquad \text{[Okapi TF]}$$

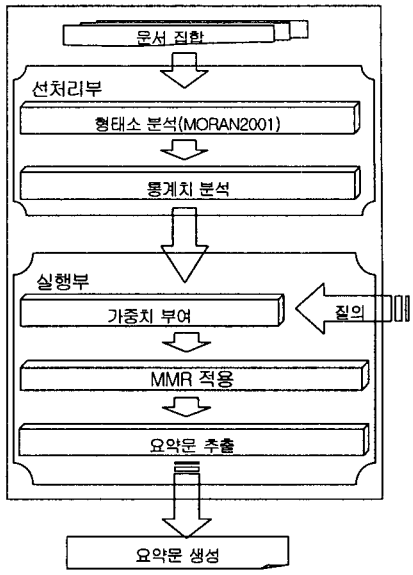
Passage안의 색인어 수가 다를때는 색인어의 수가 많은
문장이 더 유리할 수 있다. 따라서, Passage안의 색인어 수로 가중치를 나누어 정규화 한다.

$$W_i = \frac{\sum_{j=1}^{|passage|} (TF_{ij} \times idf(w_{ij}))}{|Passage|}$$

[가중치 부여 수식]

본 논문에서는 기본적으로 위 수식을 사용한다. [5] 단, Passage안의 색인어의 수를 고정하였으므로 정규화가 필요치 않으나 가중치의 범위를 제한하기 위하여 Passage상수를 사용한다.

MMR은 주어진 문장 집합(요약문)에서 유사한 문장들을 제거하는 기법이다. [6] 알고리즘은 다음과 같다.



[그림1] 질의 기반 문서 자동요약 시스템의 구조

1. 가중치로 문장들에 랭킹을 부여한다.
2. Threshold 수만큼 상위랭킹의 문장을 선택리스트에 추가한다. (본 논문에서는 Threshold=1)
3. 선택되지 않은 문장과 선택리스트의 문장과 유사도를 계산한다.
4. MMR수식에 의해 (가중치 - 유사도)값이 가장 큰 문장을
   선택리스트에 추가한다.
5 생성할 요약문장의 수만큼 3, 4, 5를 반복한다.

$$MMR = \arg MAX_{Si \in D \setminus S}[W(S) - MAX_{Sj \in S} Sim(S_i, S_j)]$$

[MMR 수식]

D\S는 선택되지 않은 문장 집합이며, S는 선택리스트 문장 집합이다. Si는 D\S의 선택되지 않은 문장이며, Sj는 선택리스트의 문장이다. Si와 Sj의 유사도 비교는 Cosine 유사도를 사용하였다. 질의를 고려하였을 때 W(Si)는 maxSim(Si, Query)가 되나, 본 논문에서는 실험문서집합과의 비교를 위하여 질의를 고려치 않은 Si의 가중치로 하였다.

이 수식은 Carbonell[6]이 제안한 수식을 문서요약에 맞게 변형시킨 것이다.

4. MMR을 적용한 시스템의 실험결과

본 논문에서는 문서 자동요약 시스템이 생성한 요약문의 성능을 알아보기 위하여 실험 문서집합에서 저자가 작성한 요약문과 유사도 비교실험을 하였다.

유사도 비교는 벡터모델을 이용한다. 논문의 저자가 중요하다고 생각한 단어는 요약문에 있을 것이라는 가정하에 논문의 저자가 작성한 요약문(이후 Abstract)과 시스템이 생성한 요약문(이후 Summary) 사이의 동일한 색인어의 수를 유사도로 사용한다. 각각의 요약문에서 색인어를 추출한 후 빈도수로 정렬하여 상위랭킹10위 안의 단어간 동일한 단어의 수를 계산한다.

MMR을 적용하지 않은 시스템에서 적용한 가중치 수식은 [4]를 참조하였다. [표1] 이중 이진과 단순 가중치 수식은 그 의미가 적어 제외하였다.

수식 이름	수식
이진	$TF = 1(if\ tf > 0),\ 0$
단순	$TF = tf$
Log	$TF = 1 + \log(tf)$
DoubleLog	$TF = 1 + \log(1 + \log(tf))$
Root	$TF = \sqrt{tf}$
Okapi	$TF = \frac{tf}{2 + tf}$
DoubleLog2	$TF = 1 + \log_2(1 + \log_2(tf))$
루트직선	$TF = \frac{tf + 3}{4}$

[표1] 적용된 가중치 수식

[표2]는 MMR을 적용하지 않은 시스템과 적용한 시스템의 유사도를 비교한 것이다.

결과에서 알 수 있듯이 DoubleLog와 Okapi의 성능 향상이 높고, 루트직선은 MMR을 적용하지 않은 시스템에서 유사도가 가장 높지만 MMR을 적용하였을 때 유사도의 변화가 없다. 이는 루트직선이 가중치에서 차지하는 TF의 비중이 높아 MMR의 영향이 거의 미치지 못하기 때문이다. 이와 반대로 DoubleLog와 Okapi 수식은 TF의 비중과 MMR의 비중이 비슷하여 MMR을 적용하였을 때 가장 성능이 좋다.

	기존 시스템	MMR	증가율
Okapi	2.21	2.40	9%
Log	2.45	2.56	4%
DoubleLog	2.10	2.31	10%
DoubleLog2	2.31	2.45	6%
Root	2.76	2.75	-0.3%
루트직선	3.33	3.33	0%

[표2] MMR을 적용하지 않은 시스템과 적용한 시스템의 비교실험

[그림2]는 MMR을 적용하지 않은 Summary 와 MMR을 적용한 Summary이다.

**MMR을 적용하지 않고 생성한 Summary

<DOC#4>
<1>결과는 IM-E는 평균 4.60초이고 IM-K는 4.48초이다. IM-E가 IM-K보다 성능이 저조한
<2>IM-E는 0.34초, IM-K는 0.41초이다. IM-E가 IM-K보다 우수
한 성능을 보이는 이유는
<3>소요되는 시간은 IM-E가 39.30초이고 IM-K가 49.30초이다. IM-E가 IM-K보다 성능이 좋은 이유는
<4>IM-K는 평균 5.25초이다. IM-E가 IM-K보다 우수한 이유는
특정 엘리먼트가
<5>시간이 6.44이고 IM-K는 11.84초가 소요된다. 이때, IM-E가
IM-K보다 성능이 우수한 이유는

**MMR을 함께 적용하여 생성한 Summary

<DOC#4>
<1>문서 1건당 소요되는 시간은 IM-E가 39.30초이고 IM-K가 49.30초이다. IM-E가 IM-K보다 성능이 좋은 이유는
<2>K-ary 완전 노드 넘버와 문서 넘버로 구성된 복합키로서 B+
트리를
<3>IM-E라 호칭한다)를 기존의 K-ary 기반 인덱스 관리자 (Index Mangaer based
<4>요소에 해당하는 엘리먼트가 없을 경우, IM-K는 계산된 K-ary 완전 트리 노드 넘버를
<5>SCL(siple Concordance list)[2]와 SERI와 전북대의 공동연구
의 K-ary 트리 방법[3]이 있다. 그러나 둘다 역화일

[그림2] MMR을 적용한 시스템과 적용하지 않은 시스템의 Summary

5. 결론

[표2]를 통해 MMR을 적용한 시스템의 성능이 높다는 것을 알 수 있었다. 이를 증명하기 위해 저자가 작성한 Abstract와 유사도 비교 실험을 하였으나 더욱 많은 문서 집합에 대해서 비교 실험이 필요하겠다.

또한, 그림2에서 볼수 있듯이 Passage 단위로 하였을 때, 문장의 길이를 선택적으로 할 수 있다는 것과 고정길이로 시스템의 구현이 쉽다는 장점이 있으나 결과 Summary의 가독성이 떨어진다는 단점이 있다. 따라서, 이 후에는 이를 보완할 수 있도록 문서의 원래 문장을 인식할 수 있는 시스템이 필요하겠다.

참고 문헌

[1] 이유리, 최기선 "수사구조를 이용한 텍스트 자동요약" 제11회 한글 및 한국어 정보처리 학술대회

1999
[2] 한경수 "질의분해를 이용한 적합성 피드백 기반 자동 문서요약", 고려대학교 컴퓨터학과 석사학위논문, 2000.
[3] 최유경, 안동언, 정성종 "다중 스레드를 지원하는 한국어 형태소 해석기" 제13회 한글 및 한국어 정보처리 학술대회 2001.
[4] 이재윤, 최보영, 정영미 "문헌 자동분류에서 용어 가중치 기법에 대한 연구" 제7회 한국정보관리학회 학술대회 논문집, 41-44, 2000.
[5] 김금영, 강인호, 안동언, 정성종, 박순철 "질의기반 자동 문서요약" 제17회 한국정보처리학회 학술대회 2002.
[6] Jaime Carbonell, Jade Golstein "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries" In Proceedings of SIGIR 1998.
[7] K. Sparck Jones "Automatic Summarizing: Factors and Directions" Advances in Automatic Text Summarization, MIT Press, pp. 1-12 1999.