문서의 인위적 요약과 통계적 알고리즘의 비교 및 분석

김유식, 유준현, 박순철
전북대학교 정보통신학과
전화 : 063-210-9351 / 핸드폰 : 018-503-5862

Comparison and analysis of artificial summary and statistical algorithm of document

Yu Sik Kim, Jun Hyun Lyu, Soon Cheol Park
Dept. of Electronic & Information Engineering, Chonbuk University
E-mail : yusikkk@cein.or.kr

Abstract

Today with the sheep of information which is produced the variety is increasing geometrical progression. To recently the internet being supplied quickly, will reach and the computer users whom it uses increase and the documents which have become digital anger are increasing. From the dissertation which it sees directness it extracts a weight with possibility work and it uses it summarizes a statistics algorithm technique and a sentence. The summary literature course which the summary and the person due to a statistics algorithm summarize an agreement ratio it compares and it compares. And being more accurate like this statistical base summary method more little more, the good hit rate is high and it proposes the document summary algorithm method which is good.

I. 서론

정보는 사람이 창조된 이후부터 시작되었다고 해도 과언이 아닐 것이다. 시대가 진행될수록 생산되는 정보의 양과 다양성은 기하급수적으로 늘어나고 있으며 최근에는 인터넷이 급속히 보급되고 이를 이용하는 컴퓨터 사용자들이 늘면서 디지털화된 문서들이 증가하고 있다. 예전에 주로 종이로 문서를 처리하던 회사와 관공서들도 점차적으로 전자적인 시스템으로 바뀌고 있다. 인터넷상의 정보량의 폭발적 증가, 컴퓨터 하드웨어의 고성능화, 문서자료의 대용량화 등은 문서의 접근 측면에서 문서요약에 대한 관심을 증가시켰다. 자동 문서요약은 몇 가지 단점에도 불구하고 비용과 효율성 측면에서 수작업에 비해 장점을 가지고 있다.

본 논문에서는 가중치를 수작업으로 직접 추출하고 통계학적 알고리즘기법을 이용하여 문장을 요약한다. 이 통계학적 알고리즘에 의한 요약과 사람이 요약한 요약문과 비교하여 일치율을 비교한다. 그리고 이러한 통계기반 요약 방법보다 좀더 정확하고 적중률이 높은 좋은 문서 요약 알고리즘 방법을 제안한다. 사람의 요약 방법에 기초한 다른 새로운 통계학적 요약 방법을 제안하고 이 방법이 사람이 요약한 요약과 얼마나 유사한지를 평가한다.

II. 문서 요약 관련 연구

지능적이고 오류가 적은 전자적 문서요약은 정보의 양이 많아지고 그것을 빠르게 찾아보려는 사람들의 욕구가 커짐에 따라 점차 중요성이 증대되고 있다. 이에 따라 요약방법이 여러 가지 제안되고 있다.

2.1 문서 요약의 여러 가지 유형

질의를 부여하는 특정 사용자의 관심 사항에 중점을 두고 요약을 제시하는 형태를 질의 기반 요약이라 하며, 질의에 상관없이 문서를 요약하는 것을 포괄적 요약이라 한다.

내용을 요약하고 축약하는 방법이라기 보다는 단순히 원문이 어떤 내용에 관한 것인지 만을 제시하는 것을 지시적 요약이라 한다. 문서 하나에 대한 요약 작업을 수행하는 것을 단일 문서 요약이라 하며, 여러 문서를 하나의 요약문에 표현하는 것을 다중 문서 요약이라 한다.

객관적인 입장을 견지하고자 하는 요약을 중립적 요약이라 하며 어떤 요구되는 관점에서 문서 내용을 추출하고 형식화하는 것을 편향적 요약이라 한다.[1][3][4]

2.2 요약 시스템의 분류

문서 요약은 문자가 생성되고서부터 이뤄져 왔다. 문서요약을 자동화하기 시작한 때는 1960년대 중반부터이다. 이때부터 문서 요약 시스템이 개발되고 연구되었으나 그 수준이 미미하다가 1990년대 중반에 접어들어서야 인터넷의 폭발적 증가와 더불어 인터넷 문서의 증가, 문서의 디지털화가 진행되면서 문서요약의 자동화 시스템이 점차 떠오르고 있는 것이다[2].

요약 시스템은 기준에 따라 여러 가지로 분류할 수 있다. 문장추출 기반 시스템, 문장 이해 기반 시스템, 혼합된 형태의 시스템, 틀(template) 기반 시스템 등이 있다.[2]

Ⅲ. 시스템에 사용한 문서 요약 기법

3.1 통계정보 알고리즘

통계정보 알고리즘이라는 것은 문장의 가중치를 계산하고, 계산된 문장의 가중치중 높은 것을 선택하여 문장의 우선 순위를 정하는 알고리즘이다. 문장의 가중치를 계산해 내기 위해 구성되는 요소는 일반적으로 세 가지를 구성한다.[3] 문장의 가중치를 계산하기 위하여서 문장 내 단어의 가중치를 계산한다. 구성되어있는 단어들의 가중치를 더하고 문장길이 정규화를 거치면 문장의 가중치가 계산된다. 중요성이 인정된 문서들을 추출하여 요약문을 생성하는 것이다.

$$\arg\max_k \frac{\sum_{i=1}^{|sentence|} f_{ij}}{|sentence|} \quad (1)$$

$f_{ij}=j$ 번째 문장의 i 번째단어의빈도수

$|sentence|$:문장에사용된단어수

단어수로 나누어주는 것은 문장마다 문장을 구성하고 있는 단어의 개수가 모두 틀릴 것이기에 문장 길이에 따른 값을 보정하기 위한 것이다.

3.2 인위적 요약 특성

인위적 요약(사람에 의한 요약)은 모든 줄거리를 머리에 생각하고 있는 상태에서 전체 내용을 대표할 수 있는 다섯 문장을 선택하여 각 문서마다 5문장씩 요약문으로 선택하였다. 이것이 다음 표1과 같다.

표1 인위적 요약과 통계적 알고리즘에 의한 요약 비교

문서	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
통계적 요약	3,5,7,9,12	2,3,8,9,11	2,3,5,9,10	2,3,7,10,11	1,3,5,7,8	1,2,3,8,7	1,3,8,9,10	1,4,5,8,7	2,4,5,9,12	3,4,5,8,7	1,3,5,7,8	2,3,8,10,10	1,3,8,9,10	1,2,3,8,9	1,4,7,8,11	3,7,8,9,11	1,2,5,7,10	2,4,5,10,12	1,4,7,8,10	4,8,10,11,12
사람의 요약	1,2,4,5,9	1,2,3,6,11	1,2,3,9	1,2,3,8,13	1,2,4,8,10	1,2,3,4,7	1,2,3,4,7	1,3,4,7,9	1,2,4,6,7	1,2,3,4,8	2,3,4,6,7	1,2,3,5,7	1,3,4,6,9	1,3,4,7,9	1,3,5,6,8	1,2,4,7,8	1,2,3,7,8	2,3,4,5,5	1,2,4,5,7	1,3,4,6,11

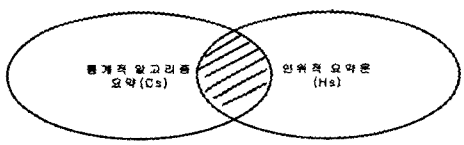
표1에서 문서번호는 문서 20개의 번호를 말하며, 각각 요약된 다섯 개의 번호는 그 문서에서 부여된 문장번호이다.

3.3 새로운 통계정보 알고리즘 요구

본 연구에서 요약의 비교를 위해 *FScore*를 사용한다. *FScore*는 인위적 요약문과 통계정보 알고리즘 요약문과 일치한 문장수를 선택되어야 할 요약문장수으로 나눈 값이다. 식으로 나타내면 다음과 같다.

$$FScore = (Cs \cap Hs)/\text{요약문장수} \quad (3)$$

(단, Cs는 통계적 알고리즘 요약문,
Hs는 인위적 요약문)



통계정보 알고리즘에 의한 요약과 인위적 요약의 결과를 비교하면 표1에서 51개 문장이 일치하여 *FScore*가 0.51을 보이고 있다. 이는 단순한 기존의 통계정보 알고리즘에 의한 요약은 사람이 원하는 바에 적중하는 확률이 높지 않음을 말하며 동시에 이는 수정되어야 할 소지가 있음을 알 수 있다. 새로운 통계정보 알고리즘을 위한 가중치가 요구된다.

Ⅳ. 실험결과

요약시스템 성능 평가에 사용되는 문서 집단에 따라서 값이 달라질 수 있기에 문서 집단을 선정하는 것은 매우 중요한 일이다.

이 장에서는 문서의 출처와 기존 통계학적 알고리즘에 의한 요약과 사람이 요약한 요약문과 비교하여 새로운 가중치가 왜 요구되는지를 밝히고 그 새로운 가중치를 구하는 공식을 생성하고 적용하여 실험할 것이다.

4.1 실험 데이터

4.1.1 실험에 사용된 문서와 가중치

본 논문의 실험에서 사용한 데이터는 한국일보(korea times) 신문 기사들로 2000년 10월부터 2001년 4월까지 나온 기사 중 국내 뉴스 20건을 뽑아서 사용하였다. 전체 문장 개수는 229개이고 전체 단어수는 2234개이다. 평균 문장 개수는 11.45개, 문장 당 평균 단어수는 9.76개이다.

아래의 그림은 사용자들이 주로 몇 번째 문장에 높은 점수를 주었는지를 나타내고 있다. 주로 앞에 있는 문장들이 중요한 문장들로 선택되었음을 보여주고 있다. 이는 시사성 뉴스의 특성상 중요한 문장이 앞에 위치하기 때문인 것으로 생각된다. 즉, 앞 부분의 문장에 더 높은 가중치를 주어야 한다는 요구이다.

표2 인위적 요약의 문장 번호와 횟수

순서 / 문장번호	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	합계
1	89	69	69	73	84	100	76	81	97	88	40	64	48	57	63	58	62	69	72	53	1412
2	63	70	56	67	76	71	60	64	32	47	61	34	18	42	38	32	53	45	42	29	1000
3	33	64	51	50	51	51	65	47	31	65	34	55	43	52	41	57	56	33	41	43	963
4	68	29	58	39	25	73	54	72	60	75	40	33	59	43	43	27	46	30	45	40	959
5	67	23	41	34	31	71	42	26	32	37	25	42	30	22	20	23	46	54	52	37	755
6	20	52	13	25	42	45	30	17	33	42	51	23	41	26	36	30	22	28	29	42	647
7	22	44	14	25	44	61	45	50	52	22	44	37	33	48	24	35	20	45	48	27	740
8	18	32	53	46	58		21	25	17	11	20	37	27	32	25	32	45	25	32	28	584
9	53	11	63	37	40		19	55	26	33	8	17	35	27	12	22	33	33	29	20	573
10	19	12	27	20	49		34	35	13	22	26	18	19	27	33	30	23	28	15	26	476
11	15	84	45	19			46	28	21	23	13	29	43	24	21	16	32	18	34	36	547
12	33	10		10				0	34	27	32				13	20		29	6	40	254
13				55											25			15			95
5개 택 장	1,4,9	11,3,6	1,9,3	1,2,3,8	1,2,10	1,4,3	1,3,7	1,4,3	1,4,12	1,4,6	2,6,3	1,3,2	4,1,9	1,3,8	1,4,6	1,3,6	1,3,8	1,5,4	1,5,2	1,3,11	

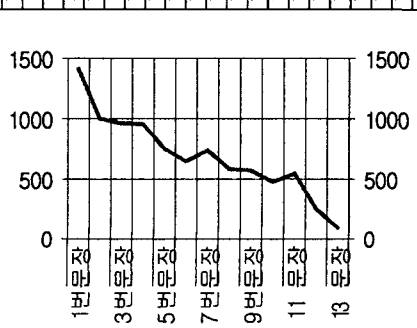


그림1 인위적 요약에 있어서 문장번호와의 상관관계

실제 기존의 통계학적 알고리즘에 의한 요약문과 사람에 의한 요약문을 비교하면 일치한 요약문장이 51개이다. *FScore*가 0.51로 낮은 편이다.

문장에서 때로는 단어가 아주 적게 쓰였음에도 불구하고 전체의 요약문장으로 나타나는 경우도 있다. 그래서 한 문장 구성이 5단어 이하인 경우를 배제한 경우와 6단어 이하 문장을 배제한 경우를 각각 사람이 선택한 것(인위적 요약)과 비교하였다. 일치한 문장수가 57개와 58개이다.

이 실험을 통해서 *FScore*가 기존 통계적 알고리즘(*Wold*)에 의한 요약문에서는 0.51, 5단어 이하 문장을 배재한 요약에서는 0.57, 6단어 이하인 문장을 배제한 요약에서는 0.58로 *FScore*가 증대됨을 알 수 있다. 이는 한 문장에서 너무 적은 단어(본 연구에서는 5단어, 6단어 이하 문장)를 사용할 때 요약문으로 채택이 어렵다는 것을 말한다.

이러한 여러 접근 방법을 적용하였음에도 불구하고 아직도 문서 요약에 기존의 통계적 알고리즘에 의한 요약은 인위적 요약문과의 *FScore*가 0.6을 넘지 못함을 알 수 있다. 어떤 결과도 사람이 선택한 것과 60%를 일치시키지 못함을 알 수 있다.

4.1.2 인위적 요약은 앞 문장에 비중이 크다(새로운 가중치 공식)

그림1을 보면 대부분의 문서가 두괄식에 의해 쓰여져 있음을 알 수 있다.

두괄식문에 대한 가중치 배려를 위한 새로운 가중치를 생성하는 식은 다음과 같다.

((총문장수)/문장번호)×기존 통계적 알고리즘 가중치

식(2)에 의한 가중치를 부여하는 것이 효과적임을 연구를 통하여 알게 되었다. 여기서 기존 통계적 알고리즘 가중치 공식(1)을 *Wold* 라고 하고, 새로운 가중치 공식을 *Wnew* 라고 한다. 식으로 나타내면 다음과 같다.

$$Wold = \arg\max_{k} \frac{\sum_{i=1}^{|sentence|} f_{ij}}{|sentence|} \quad (1)$$

$$W_{new} = W_{old} \times \left(\frac{\text{총문장수}}{\text{문장번호}} \times \alpha \right) \quad (2)$$

4.2 결 과

새로운 식(2) 가중치에 의한 요약문과 사람이 요약한 요약문을 비교하여 일치한 결과를 표3에 보인다.

표3 새로운 가중치에 의한 요약과 비교

문서	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
새가중치요약	1,2,3,5,7	1,2,3,4,8	1,2,3,7	1,2,3,5,7	1,2,3,4,8	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4,6	1,2,3,4,5	1,2,3,4,7	1,3,4,5,8	1,2,3,4,6	1,2,3,4,5	1,2,3,4,7	1,2,3,4,5	1,2,3,4,5	1,3,4,5,7	1,2,3,4,6
사람이요약	1,2,4,5,9	1,2,3,6,11	1,2,3,4,9	1,2,3,8,13	1,2,4,6,10	1,2,3,4,7	1,2,3,4,7	1,3,4,7,9	1,2,4,7,12	1,2,3,4,6	2,3,4,6,7	1,2,3,5,7	1,3,4,6,9	1,3,4,7,8	1,2,3,4,6	1,3,6,7,8	1,2,3,4,8	1,2,3,4,5	1,2,4,5,7	1,3,4,6,11
일치	1,2,5	1,2,3,6	1,2,3,4	1,2,3	1,2,4,6	1,2,3,4	1,2,3,4	1,2,3,4	4	1,2,3,4	2,3,4,7	1,2,3,5,7	1,3,4,6	1,3,4	1,2,3,4	1,3,7	1,2,3,4	1,2,3,4,5	1,2,4,5,7	1,3,4,6
일치률	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	100	100	80

일치한 문장 수가 77개로 *FScore*가 0.77이나 된다. 이는 놀라운 적중률이다. 기존 통계 알고리즘보다 훨씬 적중률이 높은 알고리즘임이 증명된다.

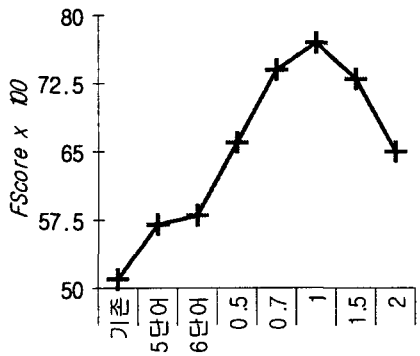
4.3 새로운 가중치 공식(*Wnew*) 적용한 성능 측정

문장번호에 새로운 가중치를 부여하는 식(*Wnew*)은 식(2)이다. 기존 통계 알고리즘에 의한 요약과 새 가중치 연구의 방법에 따른 결과 값을 표와 그림으로 나타내면 다음과 같다.

표4 각 요약의 *FScore* 비교

순	요약 방법	*FScore*	비고
1	통계 알고리즘 적용한 요약	0.51	*Wold* 적용
2	5단어 이하 문장 배제 요약	0.57	
3	6단어 이하 문장 배제 요약	0.58	
4	α= 0.5 적용하여 요약	0.66	*Wnew* 적용
5	α= 0.7 적용하여 요약	0.74	*Wnew* 적용
6	α= 1 적용하여 요약	0.77	*Wnew* 적용
7	α= 1.5 적용하여 요약	0.73	*Wnew* 적용
8	α= 2 적용하여 요약	0.65	*Wnew* 적용

그림2 각 요약 FScore



V. 결론

본 논문에서 제안하는 알고리즘은 기존 문서 요약 방법에 대해서 알아보고, 요약할 때 통계학적 알고리즘을 적용한 것과 인위적 요약(사람에 의한 요약) 방법을 비교하며 더 정확하고 사람이 원하는 요약을 자동으로 구현할 수 있는 가중치 적용에 대한 연구이다. 문서에 대한 정확한 전달에 큰 역할을 수행하고 있는 문서 요약에 대해서 알아보고, 요약을 할 때에 어떤 기준으로 문장을 요약할 것인지에 대한 문장 가중치를 설정하는 알고리즘 구현 방법이다. 문서 자동 요약 시스템에 적용하면 자동으로 문서를 요약할 때 유용하게 사용되는 알고리즘을 제안하였다.

문서 요약에는 여러 유형이 있고, 시스템 구현 방법에도 여러 가지가 존재한다. 본 논문에서는 문서의 통계적인 정보를 사람이 하는 방식과 어떻게 하면 동일한 결과를 얻을 수 있을까?하는 차원의 알고리즘을 개발하였다. 문서의 통계적인 정보를 이용하면 별다른 자료가 필요 없이 문서내의 정보를 이용하므로 쉽게 접근이 가능하다. 그러나 문서는 모두가 같은 종류의 것도 아니고 모든 문서를 단순한 가중치에 의존해서 요약한다는 것은 석연치 않다. 문장의 가중치를 계산할 때 반드시 사람의 심리적, 언어적 이해와 상황이 고려되어야 한다는 것이다. 특히 인터넷에 공개된 문서나 시사성이 있는 뉴스 같은 문서는 더욱 그러하여 문서의 앞 부분에 중요 내용을 기술한 경우가 태반이다. 그래서 단어의 빈도수에 따른 단순한 가중치(*Wold*)로는 원하는 요약을 할 수 없다는 것이다.

첫째로 한 문장을 구성하는 단어의 수가 너무 적을 때에 전체를 대표하는 요약문으로 선택이 어렵다는 것을 보여준다. 그 단어의 수는 다섯 문장이하의 것과 여섯 단어 이하의 문장이다. 이를 배제한 경우가 기존의 가중치공식을 적용한 것보다 약간씩(6%와 7%) 올라감을 알 수 있다.

둘째로 이러한 새로운 알고리즘(*Wnew*)에 대한 대두가 요구됨에 주목하여 새롭게 문장 번호에 가중치를 부여하는 방식을 취했다. 요약의 *FScore*는 0.77이다[표 4]. 기존 통계적 방법(*Wold*)에 의한 *FScore*는 0.51이었으므로 26%나 높은 훨씬 정확한 *FScore*를 보인다. 수치상으로도 어느 정도 괜찮은 시스템임을 알 수가 있다.

VI. 참고문헌

[1] 김영택 외 공저, 자연언어처리, 생능출판사, 2001.9.

[2] 강상배, 한국어문서의 통계적정보를 이용한 문서요약시스템 구현, 부산대학교, 전자계산학과 , 석사 학위논문, 1998. 2.

[3] 한경수, 질의분해를 이용한 적합성 피드백기반자동문서요약, 고려대학교 컴퓨터과학과 석사학위논문, 2000.

[4] J. Kupiec, J. Pedersen, and F.Chen, "A Trainable Document Summarizer," Proc. of 18-th SIGIR Conference, pp68-73, 1995