사단법인 한국정보과학회

인간과 컴퓨터 상호작용 연구회 회보 6권 1호

HCI '97 학술대회

발표 논문집

1997년 2월 18일(화) - 2월 20일(목)

피닉스 파크 컨벤션 센터

- 주최 : 한국정보과학회 HCI 연구회
- 주관 : 서강대학교 산업기술연구소
- 후원 : 한국과학기술원 인공지능 연구센터, (주)솔빛

Relevance Feedback을 이용한 정보검색시스템의 검색 효율 향상[1]

박 세 진*, 강 상 배**, 권 혁 철**
*부산대학교 인지과학협동과정, **부산대학교 전자계산학과

Improving the Efficiency of Information Retrieval System Using Relevance Feedback

Se-Jin Park*, Sang-Bae Kang**, Hyuk-Chul Kwon**
*Interdisciplinary Research Program of Cognitive Science, Pusan National University
**Dept. of Computer Science, Pusan National University

요약

검색 비전문가가 질의문을 정확하게 작성하여 원하는 문서를 검색하기는 매우 어렵다. 이 논문은 이 문제를 해결하기 위하여 Relevance Feedback 방법을 사용한다. Relevance Feedback 은 검색한 문서 중에서 적합하다고 판단한 문서에 있는 색인어를 질의어에 추가하여 다시 검색하는 방법이다.
실험방법은 5 가지의 질의문 수정 방법에 따른 검색 효율을 비교한다. 실험에 사용하는 정보검색시스템은 부산대학교 인공지능연구실에서 개발하였으며, 실험데이터는 부산일보 신문기사 5만 3천 건과 KT-Set 2.0 이다.

서 론

정보의 양이 폭발적으로 증가함에 따라 정보검색시스템(Information Retrieval System)의 필요성이 증가하였다. 하지만 정보검색 훈련을 거의 받지 못한 일반 이용자(end-user)는 정보검색시스템을 이용하여 원하는 정보를 효과적으로 찾기가 어렵다. 그 이유는 정보검색시스템의 색인어와 이용자의 탐색어가 다르고, 이용자는 자신이 원하는 정보를 정확한 용어를 표현하기 어렵기 때문에 최적의 질의문을 작성하지 못한다. 정확하지 않은 질의문으로는 원하는 정보를 검색할 수 없다. 그래서 대부분의 이용자는 보다 나은 검색결과를 얻기 위해 탐색을 반복한다. 이용자는 질의문을 수정하여 다시 검색하는데, 이용자는 검색 결과에 만족할 때까지 질의문 수정을 반복한다. 따라서 정보 검색시스템은 이런 질의문 수정을 자동으로 해줌으로써 사용자가 사용하기에 편리한 환경을 제공해 주어야 한다.

이 논문은 질의문을 수정하여 더 많은 적합 문서를 검색하기 위해 적합성 피드백(relevance feedback) 방법을 사용한다. 적합성 피드백[3, 4, 5, 6] 은 검색한 문서 중에서 적합하다고 판단한 문서에 존재하는 색인어를 질의문에 추가하여 다시 검색하는 방법이다. 적합 문서에 나오는 색인어들은 원질의어와 관련성을 가진다. 이런 관련성 정보를 이용하여 원래의 질의문에 색인어를 첨가하면 더 많은 적합 문서를 검색할 수 있다. 그러나, 관련성이 적거나 또는 전혀 관련성이 없는 색인어를 원질의어에 추가한다면 1 차 검색보다 더 적은 수의 적합 문서를 검색한다. 결국, 적합 문서에 나오는 색인어 중에서 어떤 색인어를 선택하여 원질의문에 추가할 것인가에 대한 문제가 발생한다. 이 논문은 이용자가 적합하다고 판단한 문서에서 질의문을 수정

본 연구는 '96 년도 제 1 차 한국과학재단 산학협력연구과제 (96-2-11-02-01-3) 연구비 지원을 받고 있음.

할 색인어를 선택하는 방법에 따른 검색효율의 차이를 비교한다. 부산대학교 인공지능연구실에서 개발한 시스템을 이용하여 실험한다.

먼저 적합성 피드백 방법을 설명하고, 질의문을 수정할 색인어를 선택하는 방법을 설명하고, 실험 결과를 알아보고, 마지막으로 결론과 앞으로의 연구방향을 설명한다.

2. Relevance Feedback

정보검색시스템은 질의문의 질의어와 검색 대상 문서의 색인어를 유사도 함수를 이용하여 비교한다. 질의어와 높은 유사도를 나타내는 문서일수록 순위가 높다. 이용자는 일차 검색 문서 중에서 적합한 문서와 적합하지 않은 문서를 판단한다. 이용자가 관련이 있다고 판단한 문서의 색인어는 원질의어와 관련성을 가진다. 이 관련성에 기반하여, 적합 문서내에 존재하는 색인어를 원질의어에 첨가하여 재검색을 하면 적합하다고 판단한 문서의 주제와 유사한 주제를 가진 문서가 찾아질 가능성이 높다.

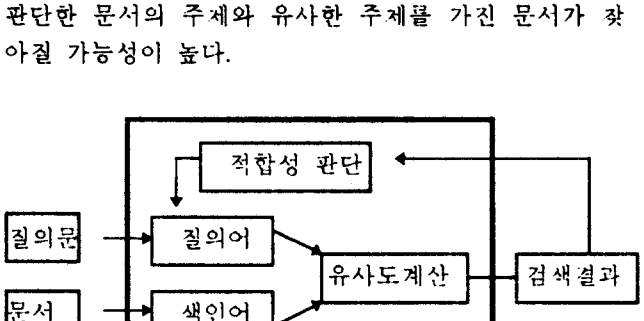


그림 적합성 피드백

그러나 그 반대로 적합 문서내에 존재하는 색인어 중 불용어 또는 빈도가 낮은 색인어는 원질의어와 직접적인 관련이 없음에도 불구하고 질의어에 첨가되므로, 검색 문서 수가 증가하고 관련이 없는 문서가 검색되는 단점이 있다. 따라서 적합 문서에 존재하는 색인어 중에서 원질의어에 첨가되는 색인어는 신중하게 선택해야 하며, 색인어 선택방법에 따라 검색효율에 많은 차이가 발생한다. 따라서 본 논문은 질의문을 수정할 색인어를 선택하는 방법에 따른 검색 효율을 비교한다.

3 색인어 선택 방법

적합성 피드백에서 원질의문에 첨가할 색인어를 선택하는 것은 대단히 중요하다. 특히, 적절치 못한 색인어의 선택은 사용자가 전혀 원하지 않는 검색 결과를 낳기 때문이다. 따라서, 색인어의 선택은 신중하여야 하며, 사용자가 입력한 질의문과 관련성을 가지는 색인어를 선택하여야 한다. 이 논문에서는 5 가지 색인어 선택방법을 사용하였다.
질의문을 수정할 색인어를 선택할 때 정보 검색 시스템이 자동으로 색인어를 선택할 수도 있고, 적합 문서에 있는 색인어 목록 중에서 이용자가 색인어를 선택할 수도 있다.

A. COM : 이용자가 적합하다고 판단한 문서에서 공통으로 나타나는 색인어는 그 문서에서 중요한 색인어일 가능성이 크다. 적합 문서가 2 개 이상일 때 문서에서 공통으로 나타나는 색인어를 모두 질의문에 첨가한다.

B. HDFCOM : 적합 문서에서 공통으로 나타나는 색인어 중 질의어의 문헌빈도보다 높은 문헌빈도를 가진 색인어를 선택하여 질의문에 첨가한다.

C LDFCOM : 적합 문서에서 공통으로 나타나는 색인어 중 질의어의 문헌빈도보다 낮은 문헌빈도를 가진 색인어를 선택하여 질의문에 첨가한다.

D USERSEL : 이용자가 적합하다고 판단한 문서에 있는 색인어의 목록을 이용자에게 제시하고 이용자가 스스로 질의문을 수정할 색인어를 선택한다.

E tf * idf : 2 개 이상의 문서에 공통으로 나타나는 색인어를 tf*idf 가 높은 순으로 정렬해서 상위 5 개, 10 개, 15 개, 20 개의 색인어를 선택하여 원질의어에 첨가한다. 단, 2 개 이상의 문서에 공통으로 나타나는 색인어가 5, 10, 15, 20 개 이하일 때는 나타난 색인어 모두를 선택한다. 따라서 실제 첨가되는 색인어의 개수는 5, 10, 15, 20 보다 적을 수도 있다. tf*idf 는 단어빈도가 높고 문헌빈도가 낮을 수록 큰 값을 가진다.

4 실험 결과

4.1. 실험 환경

이 논문에서 사용할 정보검색시스템[1, 2]은 부산대학교 전자계산학과 인공지능연구실에서 개발하였다. 시스템은 색인기, 검색기, 등록기의 세 부분으로 크게 나누어진다. 색인어를 추출하기 위해 형태소 분석, 중의성 제거, 미등록어 추정 방법을 이용한다. 자연어 질의는 색인어를 추출하는 방법에 따라 질의어를 추출하고, 벡터공간모델에 의해 검색하고, 코사인 유사계수 함수를 이용하여 문서의 순위를 결정한다.

실험 데이터는 부산일보사 데이터와 KT-Set 2.0 데이터를 사용한다. 부산일보사 95년도 기사 전문은 5만 3천 건으로, 정치, 경제, 사회, 문화, 생활, 체육 분야의 기사로 용량은 90Mbyte이다. 30개의 자연어 질의문(natural query)을 가지고 실험하고, P대학교 학생들을 대상으로 질의어를 수집하였다. KT-Set 2.0은 KT논문초록, 전자신문 그리고 잡지 기사를 포함하여 4,414건이며, 질의어는 자연어 질의문 50개이고, Kt-Set에 포함된 문서는 전자와 전산분야에 관련된 내용으로 구성되어 있다.

4.2. 결과와 분석

적합성 피드백 방법은 여러 실험에서 검색효율을 향상시키는 수단임이 입증되었는데[4][5], 대략 첫번째 피드백에 의해 정확률이 40-60% 향상되며 두 번째 피드백 과정부터는 향상률이 훨씬 줄어든다[5]. 이 논문은 적합성 피드백을 이용하여 질의문을 한번 수정하였을 때 색인어 선택 방법에 따른 검색 효율을 비교한다.

불리언 질의문을 이용한 검색과는 달리 자연어 질의문을 이용한 검색은 재현율(recall)은 높고 정확도(precision)는 떨어진다. 부산일보 데이터는 가공되지 않은 원시 신문 기사 데이터로, 검색효율을 판단하는 기준으로 상위 10위, 20위, 30위 안에 나타나는 적합 문서의 개수를 이용하였다. 부산일보 데이터나 웹 문서와 같이 적합한 문서의 수를 정확히 알지 못하는 경우에 정확도-재현율 그래프를 이용하여 검색 효율을 평가할 수 없다. 또한, 정보검색시스템 이용자는 상위에 존재하는 일부 문서에만 주의를 기울이므로, 본 논문에서는 상위 30위 까지의 문서에 포함되어 있는 적합 문서의 수로써 검색 효율을 평가하였다. 1차 검색에서 상위 30위 안에 포함되어 있는 적합 문서의 개수와 적합성 피드백을 적용한 후 2차 검색에서 상위 30위안의 적합 문서의 개수를 비교한다.

KT-Set 2.0에 대한 검색효율을 평가하는 기준으로 정확도-재현율 그래프와 상위 50위 안에 포함되는 적합 문서의 개수를 비교하는 방법 두 가지 모두 사용하였다.

4.2.1. 부산일보 데이터

[표 1]은 상위 30위 안의 적합 문서의 개수를 나타낸다.

	10위	20위	30위	증가율%
1차 검색	5	9.03	11.9	
COM	6.87	11.74	16.2	36.13%
USERSEL	6.67	11.84	16.3	36.97%
HDFCOM	5.13	8.9	12.03	1.09%
LDFCOM	5	8.4	11.4	↓
tf*idf 5개	6.7	11.6	14.73	23.78%
tf*idf 10	6.97	11.64	15.67	31.68%
tf*idf 15	7.07	11.97	16.33	37.22%
tf*idf 20	7	12.27	16.53	38.9%

[표 1] 부산일보 데이터 실험 결과 - 10위, 20위, 30위 안에 나타난 적합 문서의 수

1차 검색에서 평균 4.27개의 질의어로 평균 11.9개의 적합 문서가 검색된다. 가장 높은 검색 효율을 나타낸 방법은 "tf*idf 20"방법으로, 평균 19.46개의 색인어를 원질의어에 첨가하였고 평균 16.53개의 문서를 검색하여 38.9% 검색 효율이 증가한다. "tf*idf 15" 방법은 평균 14.73개의 색인어를 원질의어에 첨가하였고 검색 효율이 37.22% 증가한다. "tf*idf"값에 의해 색인어를 선택하면 단어빈도와 역문헌빈도를 고려하여 각 문서를 대표할 수 있는 중요 색인어를 선택할 수 있다. 색인어를 5개 첨가하는 것보다 10개를 첨가했을 때 검색되는 적합 문서 수가 더 많고, 색인어 20개를 첨가하면 색인어를 10개 첨가하는 것보다 더 많은 수의 적합 문서를 찾을 수 있다.

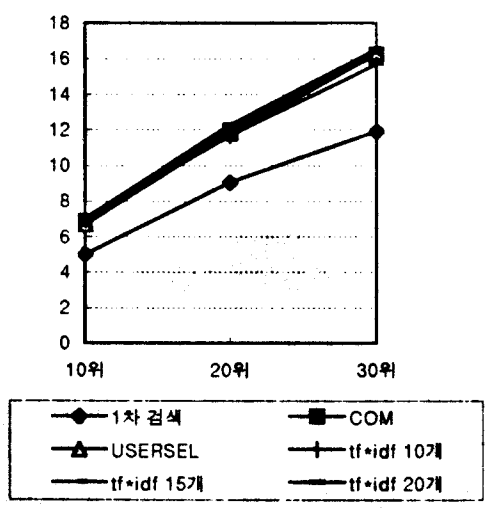
단순히 적합 문서에서 공통으로 나타나는 색인어를 추출하는 "COM"방법은 평균 11.1개의 색인어를 원질의어에 첨가하고 36.13% 검색 효율이 증가한다. 이용자가 적합 문서를 보고 원질의어에 추가할 색인어를 직접 선택하는 "USERSEL"방법은 평균 7.53개의 색인어를 원질의어에 추가하고 36.97% 검색 효율이 증가한다. "USERSEL"방법은 "COM"에 비해 훨씬 적은 수의 색인어를 원질의어에 추가하지만 비슷한 검색 효율을 나타낸다. "USERSEL"방법은 이용자가 직접 찾고자 하는 주제를 표현하기에 적합한 색인어를 선택하기 때문에 적은 수의 질의어로 많은 수의 적합 문서를 찾

을 수 있다.

"HDFCOM"은 질의어의 문헌빈도 중에서 가장 높은 문헌빈도보다 높은 문헌빈도를 가지는 색인어만을 추출하고 "LDFCOM"은 질의어의 문헌빈도보다 낮은 문헌빈도를 가진 색인어만을 추출한다. 검색된 적합 문서 수에 있어 큰 차이가 없다.

적합성 피드백을 적용하여 검색하면 적합 문서의 순위가 올라간다. 예를 들면 "바다 양식"이라는 질의문으로 1,269 개의 문서가 검색되고 상위 30 위에서 문서 16 개가 적합하다. "tf*idf 20" 색인어 선택 방법으로 원질의문을 수정해서 재검색을 하면 3,233 개의 문서가 검색되고 상위 30 위 중 문서 29 개가 적합하다. 상위 10 위 안을 보면 1 차 검색에서 상위 10 위 중에서 2, 7, 10 위의 문서가 적합하다. 적합성 피드백을 적용하면 상위 10 위 중에서 1, 2, 3, 5, 6, 7, 8, 9, 10 위의 9 개의 문서가 적합하다. 이용자는 전체 검색 문서 중에서 상위에 나타난 몇 십개의 문서에만 주의를 기울이기 때문에 상위 순위에 적합 문서가 나타난다는 것은 매우 중요하다. 11 위에서 20 위 사이에 적합 문서 4 개가 증가하는 것보다 10 위 안에 적합 문서 2 개가 증가하는 것이 훨씬 더 중요하다. 10 위 안에 평균 6.42 개, 20 위 안에 4.61 개, 30 위 안에 3.84 개로 적합 문서의 수가 증가하였다.

[그림 2]는 1 차 검색 결과와 "COM", "USERSEL", "tf*idf 10", "td*idf 15", "tf*idf 20" 색인어 선택 방법의 결과를 그래프로 나타낸 것이다. 적합성 피드백을 적용했을 때 검색 효율이 어느 정도 향상되는지 보여준다.



[그림 2] 부산일보 검색 결과

4.2.2. KT-Set 2.0

[표 2]은 상위 50 위 안의 적합 문서의 개수를 나타낸다.

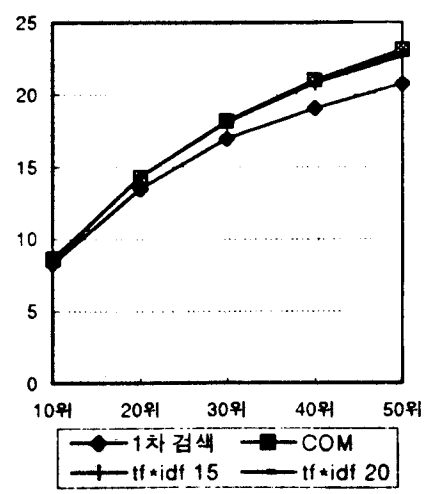
	10위	20위	30위	40위	50위
1차 검색	8.3	13.48	16.96	19.04	20.72
COM	8.6	14.30	18.2	21.02	23.12
HDFCOM	8.54	14.18	17.96	20.78	22.78
LDFCOM	8.3	13.48	16.98	19.06	20.74
tf*idf 5	8.48	13.90	17.46	19.78	21.46
tf*idf 10	8.52	14.10	17.86	20.24	21.92
tf*idf 15	8.58	14.28	18.10	20.78	22.60
tf*idf 20	8.58	14.32	18.18	20.96	22.92

[표 2] KT-Set 2.0 실험 결과

1 차 검색에서 평균 20.72 개의 적합 문서가 검색된다. 가장 높은 검색 효율을 나타낸 방법은 "COM"방법으로, 평균 6.44 개의 색인어를 원질의어에 첨가하였고 평균 23.12 개의 문서를 검색하여 11.6% 검색 효율이 증가한다. "tf*idf 20"방법은 평균 16.26 개의 색인어를 원질의어에 첨가하였고 검색 효율이 10.6% 증가한다. "tf*idf 15"방법은 평균 13.08 개의 색인어를 원질의어에 첨가하였고 검색 효율이 9% 증가한다. 그리고 첨가하는 색인어의 수가 5 개, 10, 15 개, 20 개로 많아지면 검색되는 문서의 수도 증가하지만 동시에 상위 30 위 안에 적합 문서의 수도 늘어난다.

"HDFCOM"방법은 평균 5.5 개의 색인어를 질의문에 첨가하고 10% 검색 효율이 증가한다.

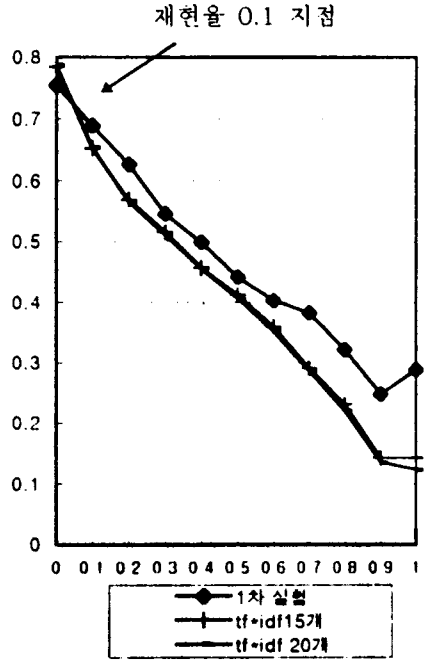
[그림 3]은 1 차 검색 결과와 "COM", "td*idf 15", "tf*idf 20" 색인어 선택 방법의 결과를 그래프로 나타낸 것이다. 적합성 피드백을 적용했을 때 검색 효율이 어느 정도 향상되는지 보여준다.



[그림 3] KT-Set 2.0

[그림 4]는 KT-Set 데이터 실험 결과의 정확도-재현율 그래프이다. 재현율 0.1 지점을 보면 적합성 피드백을 적용한 실험 결과의 정확도가 더 높지만, 재현율

이 높아질 수록 2차 검색 결과의 정확도가 1차 검색 결과의 정확도보다 더 낮아진다. 이유는 적합성 피드백을 적용한 2차 실험은 1차 실험보다 질의어의 수가 증가하여 많은 수의 문서를 검색하기때문이다. 상위 순위에는 적합문서가 많이 나타나지만 순위가 내려갈수록 부적합 문서의 수가 많이 나타나기 때문에 재현율이 높아지면 정확도가 낮아진다. 이용자는 상위 순위의 문서에만 관심을 가지기 때문에 상위 순위 안에 적합 문서가 나타나는 것이 중요하다. 그러므로 적합성 피드백을 적용하면 재현율이 높아지면 정확도는 더 떨어지지만, 적합성 피드백을 적용하면 검색 효율이 향상된다.



[그림 4] KT-Set 2.0 정확도-재현율

KT-Set 2.0은 색인어 선택 방법에 따른 검색 효율에 큰 차이가 나지 않는다. 데이터가 4000건밖에 되지 않고 각 질의문 당 평균 적합 문서의 수가 적기 때문에 적합성 피드백에 의해 추가로 검색되는 문서의 수가 적다. 이에 반해 부산일보 신문기사 데이터는 색인어 선택 방법에 따라 검색 효율에 큰 차이를 보인다. 데이터가 5만 3천건이고 신문기사라는 특성상 질의문의 주제가 일반적이다. 일반적인 주제의 질의문은 관련이 있는 문서를 많이 검색해 낼 수 있으며 이용자의 적합성 판단 기준도 낮아진다. 따라서 적합성 피드백에 의한 효과가 크다.

4. 결론

색인어 선택 방법에 의한 검색 효율을 비교해 보면 "COM", "tf*idf 20", "tf*idf 15" 색인어 선택 방법이 가장 검색 효율이 좋다. 시소러스를 사용하지 않고도 충분히 질의어를 확장하여 적합 문서를 더 많이 검색할 수 있다. 부산일보 데이터와 KT-Set 2.0의 결과를 비교해 보면 데이터의 크기가 크고, 적합 문서 수가 많을 수록 적합성 피드백의 효과가 크다.

이번 연구를 바탕으로 다양한 색인어 선택 방법을 개발하여 실험을 할 예정이다. 검색 문서 중에서 적합 문서와 부적합 문서를 판단하여 적합 문서에서 공통으로 나타나는 색인어 목록과 부적합 문서에서 공통으로 나타나는 색인어 목록을 구한다. 부적합 문서에서 추출한 색인어 중에서 적합 문서에서 추출한 색인어를 제외한 색인어를 가진 문서는 검색하지 않는다. 적합 문서의 수는 늘어나고 부적합 문서의 수는 줄어든다. 재현율과 정확도가 동시에 높아진다. 또 다른 방법은 "tf*idf"색인어 선택 방법에서 원질의어에 추가되는 색인어의 수가 많을 수록 검색되는 적합 문서의 수가 많아졌다. 원질의어에 추가하는 색인어의 수를 30개, 40개, 50개로 증가시켜 추가 실험이 필요하다. 그리고 크기가 더 큰 전문 데이터(full-text data)를 이용한 추가 실험이 필요하다.

참고문헌

[1] 이준영, 강상배, 양장모, 박 승, 박현주, 김민정, 권혁철. "다중색인에 의한 정보검색 시스템 구현", 한글 및 한국어정보처리 학술발표 논문집, pp. 63-67, 1996.

[2] 이준영. "다중색인과 압축저장에 의한 정보검색 시스템 개발에 관한 연구", 부산대학교 전자계산학과 석사학위논문, 1997.

[3] 정영미. *정보검색론*, 구미무역(주)출판부, 1993.

[4] David Haines and W.Bruce Croftt. "Relevance Feedback and Inference Networks", http://ciir.ca.umass.edu/info/psfiles/irpub/ir.html.

[5] Gerard Salton and Chris Buckley. "Improving Retrieval Performance by Relevance Feedback". *Journal of The American Society for Information Science*. 41(4):288-297, 1990.

[6] Gerard Salton and Michael J.McGill. *Introduction to Modern Information Retrieval*. McGrawHill, 1983.

[7] Jurgen Koenemann and Nicholass J. Belkin. "A Case For Interaction : A Study of Interactive Information Retrieval Behavior and Effectiveness". *CHI 96*(April 13-18). 205-

212. 1996.

[8] William B.Frakes and Ricardo Baeza-Yates. *Information Retrieval:Data Structures & Algorithms.* 1995.