논문-00-5-1-08

영상 데이터베이스 검색 시스템의 검색효율 평가를 위한 새로운 평가척도

서 창 덕*, 김 회 율*

A Novel Measure for Retrieval Efficiency of Image Database Retrieval System

Changduck Suh* and Whoi-Yul Kim*

요 약

본 논문에서는 순위가 부여되는 영상 데이터베이스 검색 시스템의 검색효율을 평가하기 위한 새로운 단일가 척도를 제안한다. 좋은 순위부여 시스템이 되기 위한 조건은 첫째, 관련영상을 많이 검색해야 하며 둘째, 부적합 영상은 검색하지 말아야 하며 셋째, 평균순위가 높아야 하고 넷째, 검색된 관련영상들이 밀집되어 있어야 한다. 기존 평가척도들이 일부 조건만을 반영하며 개략적 혹은 부정확한 평가 결과를 보이는데 반해, 제안하는 평가척도 NDS(Normalized Distance Sum)는 이러한 문제점들을 모두 해결한다. NDS의 우수성을 입증하기 위해 ${}_nC_r({}_{10}C_5=252,\ {}_{20}C_9=167{,}960)$개의 검색패턴을 자동 발생시켜 이를 기존 평가척도와 함께 측정 비교한다. 이 패턴들은 n순위 내에서 r개의 관련 영상이 검색된다고 가정하였을 때 재귀적 함수 호출에 의해 자동 발생된 것들이다.

Abstract

This paper proposes a single metric to measure and evaluate the retrieval effectiveness of image database retrieval system that requires an ordered ranking. There are four conditions to be a good ranking system. First, the number of relevant images among the retrieved should be as large as possible. Secondly, the number of irrelevant images should be smaller. Third, the average rank of relevant images should be higher. Last, the relevant images should be clustered close together. The conventional evaluation measures only reflect a part of the conditions listed above, and the evaluated results are coarse or inaccurate. The proposed NDS, however, resolves all those problems. In order to prove the efficiency of the NDS, we generate patterns of ${}_nC_r({}_{10}C_5=252,\ {}_{20}C_9=167{,}960)$ to evaluate and compare with other measures. The patterns were generated automatically by a recursive function call on the assumption that 'r' relevant images are retrieved within the range of 'n'.

I. 서 론

컴퓨터와 저장매체의 발달로 기존 문자 정보를 비롯하여 영상, 음성, 동영상 등의 멀티미디어 데이터가 급속히 증가하고 있다. 또한 인터넷의 발달로 전 세계 여러 곳에 존재하는 자료를 손쉽게 구하고 서로 공유할 수 있게 되었으며 이러한 멀티미디어 자료를 수집·가공하여 또 다른 새로운 자료를 생성하는 것이 보다 쉬워졌다. 따라서 이러한 수많은 자료를 보다 효율적으로 관리하고 검색하는 방법이 연구되어 왔다.

검색 알고리듬은 검색 매체에 따라, 또 검색 목적에 따라 달라지는데 과거 문헌검색 시절부터 현재의 내용기반 검색에 이르기까지 수많은 검색 알고리듬들이 제안되고 있다. 이러한 알고리듬에 의한 검색 시스템의 평가는 일반적으로 검색효율, 신속성, 경제성의 세 가지 측면에서 수행된다[1]. 이 중 특히 검색효율은 관련자료를 많이 검색할

* 한양대학교 전자공학과
Dept. of Electronics Engineering

수록, 그리고 부적합자료는 검색하지 않을수록 좋은 것으로, 검색 알고리듬의 성능을 정량적으로 측정하기 위해 평가척도를 이용한다. 검색효율(성능)을 측정하는 것은 그 결과에 따라 알고리듬의 우열이 나타나므로 최대한 객관적이고 세밀해야 한다.

검색 결과는 순위 없이 사용자가 원하는 내용을 무순위로 출력하는 것과, 유사도 계산에 의해 사용자가 원하는 내용에 가장 근접한 내용일수록 높은 순위에 출력하는 방법이 있다. 검색된 항목이 많아질수록 무순위로 출력하기보다는 높은 유사도를 갖는 높은 순위의 항목부터 낮은 순위의 항목 순서대로 보여주는 것이 이용자의 입장에서 보다 바람직한 형태가 된다. 특히 내용기반 검색 시스템은 이러한 순위부여 시스템인 경우가 많다.

최근 일련의 MPEG-7 표준화 회의에서 보여주듯이 자신이 제안한 알고리듬을 표준으로 삼기 위한 노력이 활발한데 이는 곧 알고리듬의 우수성을 객관적으로 입증해야 함을 뜻한다. 따라서 제안된 수많은 검색 알고리듬을 객관적으로 평가하여 표준화하기 위한 회의가 정기적으로 열리고 있다. 그러나 기존의 평가척도로는 많은 문제가 있어 본 논문에서는 검색효율을 측정하기 위한 새로운 평가척도를 제안한다.

II. 관련연구 및 문제점

데이터베이스(DB) 검색 시스템에서 검색 결과는 두 가지 형태 중 하나로 주어진다. 하나는 적합/부적합 여부에 따라 출력이 결정되며 다른 하나는 질의와의 유사도에 따라 순위가 부여되어 출력된다. 따라서 순위부여가 되지 않는 시스템을 위한 평가척도가 있으며 순위부여 시스템에 적합한 평가척도가 있다.

순위부여 시스템의 평가에는 순위정보를 이용하는 평가척도가 바람직하지만 그렇지 않은 경우도 이용된다. 그러나 순위정보를 이용하는 평가척도를 순위가 부여되지 않는 시스템에 적용할 수는 없다.

검색효율을 측정하기 위한 평가척도는 1960년대부터 제안되기 시작하였는데 그 중 과거 문헌검색 시절부터 현재까지 널리 사용되고 있는 척도로 재현율과 정확도가 있으며, 순위가 부여되는 내용기반 영상검색 시스템을 위해 비교적 최근에 제안된 평가척도로 BEP, AR, NMRR 등이 있다.

1. RR-PR

고전적인 문헌검색 효율을 나타내는 척도로는 재현율(RR: recall ratio), 정확도(PR: precision ratio), 누락율(snobbery ratio), 잡음율(noise factor), 부적합율(fallout ratio), 배제율(correct-rejection ratio)이 있다[1].

이 중 재현율(RR)과 정확도(PR)를 제외한 나머지는 현재 거의 사용되지 않는데, RR은 DB내 적합 문헌(또는 영상)들 가운데 검색된 적합 문헌들의 비율을, PR은 검색된 문헌에 대한 적합 문헌들의 비율을 말하는 것으로 수치가 1에 가까울수록 좋다.

$$RR = \frac{Rs}{R} \qquad PR = \frac{Rs}{S} \qquad \text{where } Rs: \text{검색된 적합 문헌수},\ R: \text{전체 적합 문헌수},\ S: \text{검색된 문헌수} \tag{1}$$

즉, 식 (1)은 검색된 적합문헌의 개수를 기준으로 한 척도로서 일반적으로 이들은 서로 반비례 관계에 있으며 함께 사용해야만 의미가 있다. 따라서 재현율의 변화에 따라 정확도의 변화를 그림 1처럼 2차원의 정확도-재현율 그래프(precision-recall graph)로 나타내거나 이 둘을 결합시킨 복합척도를 사용한다.

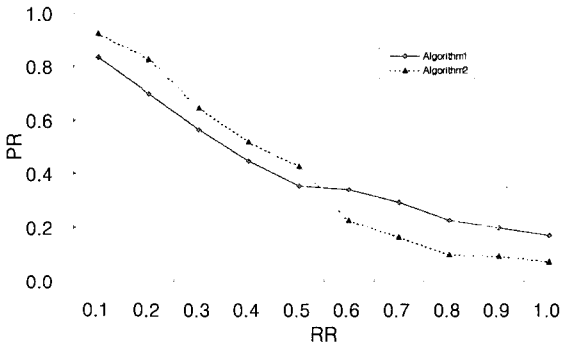

그림 2. 정확도-재현율 그래프
Fig. 1. Precision-Recall graph.

복합척도(E)는 여러 가지가 제시되어 있는데 RR과 PR을 단순히 더한 것에서부터 식 (2)처럼 이용자가 PR와 RR에 각각 상대적 가중치 α 와 β 를 설정[2]하는 등 몇 가지가 있다.

$$E = 1 - \{\alpha PR^{-1} + (1-\alpha)RR^{-1}\}^{-1} = 1 - \frac{(\beta^2+1)PR \cdot RR}{\beta^2 PR + RR} \qquad \text{where } \alpha = \frac{1}{(\beta^2+1)} \tag{2}$$

이러한 복합척도는 가중치에 따라 여러 개의 다른 값들을 갖는 문제가 있으므로 논문 [3]에서처럼 식 (2)의 β를 1로 하여 즉, RR과 PR에 같은 중요도를 부여하여 사용하기도 한다.

2. BEP류

순위부여 시스템을 위한 평가척도로 1968년에 Cooper가 제안한 예상 탐색길이 척도[4]는 PR의 성격을 가지고 있으나 반영해야하는 인자가 많고 복잡해 현재 사용되고 있지 않다. 따라서 비교적 최근인 1995년 식 (3), (4)의 평가척도가 제안되어 현재까지 사용되고 있다.

FP(Filtering Performance)는 DB 전체영상 개수의 상위 20% 순위 내에서 검색된 관련영상의 개수가 전체 관련영상 개수의 몇 %인지를 측정한 것이며, BEP(Bull's Eye Performance)는 관련영상 개수의 두 배 순위 안에 포함된 관련영상 개수를 세어 실제 관련영상의 몇 %가 검색되었는지를 측정한 것이다[5]. 이밖에 BEP의 변형된 형태로 BEP_{1R}이 있다.

$$\begin{aligned} FP(\%) &= \frac{Rs_{20}}{R} \times 100 \\ BEP(\%) &= \frac{Rs_{2R}}{R} \times 100 \\ BEP_{1R}(\%) &= \frac{Rs_{1R}}{R} \times 100 \end{aligned} \tag{3}$$

where R : 전체 관련영상 개수
Rs_{20}: DB크기의 20% 순위내 관련영상 개수
Rs_{2R}: $2R$ 순위내 관련영상 개수
Rs_{1R}: R 순위내 관련영상 개수

이들은 100에 가까울수록 좋은 검색효율을 뜻하는데 문헌검색 효율척도로 널리 사용되는 RR과 매우 유사하다. 다른 점이 있다면 식 (3)의 분자에 해당되는 검색된 관련영상의 측정범위가 제한된다는 점으로 이들은 순위부여 시스템을 위한 RR이라고 할 수 있다.

현재 MPEG-7 영상 descriptor의 Motion/Shape group에서는 제안한 알고리듬들을 평가하기 위한 척도로 BEP_{1R}과 BEP를 채택하고 있다[6]. 반면 DB의 크기가 반영되는 FP는 현재 별로 사용되고 있지 않다.

3. RP류

앞서의 BEP류는 검색된 상위 일부분을 측정범위로 하므로 순위정보를 간접적으로 이용한다고 볼 수 있지만 RP류(Ranking Performance)는 관련영상의 순위값 자체를 집적 이용한다는 점에서 앞서의 방법들과 차이가 난다.

RP류에는 모든 검색된 관련영상들의 평균 순위를 나타내는 평균순위(AR: average rank)와 마지막 관련영상의 순위를 말하는 최대순위(MR: maximum rank)가 있으며[5], 그밖에 AR의 문제점을 일부 보완한 NAR(Normalized AR)[7]이 있다. MR과 표준편차(σ) 또는 분산(σ^2)은 AR 혹은 NAR과 함께 보조적으로만 사용되며, RP류는 모두 작을수록 좋다.

$$\begin{aligned} AR &= \sum_{i=1}^{Rs} Rank(i) \;/\; Rs \\ NAR &= \sum_{i=1}^{Rs} Rank(i) \;/\; \sum_{i=1}^{Rs} i \\ MR &= \max\{Rank(i)\} \\ \sigma &= \sqrt{\frac{\sum_{i=1}^{Rs}(Rank(i) - AR)^2}{Rs}} \end{aligned} \tag{4}$$

where $Rank(i)$: 검색된 관련영상 중 i번째 순위
Rs: 검색된 관련영상의 개수

4. NRS

RP류는 [0,1]값으로 정규화되지 않기 때문에 평가값은 검색된 관련영상의 개수에 좌우된다. 그러나 1997년에 제안된 NRS(Normalized Rank Sum)[8]는 식 (5)에서 보듯이 정규화된 척도이다. NRS의 값은 1에 근접할수록 우수한 검색효율을 뜻한다.

$$NRS = \frac{R(R+1)/2}{\sum_{i=1}^{R} Rank(i)} \tag{5}$$

where R : 전체 관련영상 개수

NRS는 정규화되어 있으며 σ 나 MR을 필요로 하지 않기 때문에 모든 면에서 RP류보다 우수하다. 이 평가척도는 MPEG-7 제안서[9]에서 인용되었을 뿐 널리 알려져 있지는 않다.

5. NMRR

1999년 3월 Color/Texture descriptor group의 MPEG-7 표준화 회의에서 사용된 평가척도는 처음 AVR에서 출발

하여 그 후 MRR이 만들어졌고 수정을 거듭해 NMRR (Normalized Modified Retrieval Rank)이 만들어져 1999년 10월 회의결과 식 (6)으로 최종 정의[10]되어 추후 제안된 논문에서는 식 (6)으로 사용하고 있다. 이 척도 또한 [0,1]로 정규화 되어 있으며 0에 근접할수록 좋은 것이다.

$$NMRR = \frac{MRR}{K+0.5-0.5R} \quad (6)$$

$$where\ MRR = AVR-0.5-0.5R$$
$$AVR = \sum_{i=1}^{R} Rank(i) \ / \ R$$
$Rank(i)$: i번째 관련영상의 순위
$K+1$(i번째 관련 영상이 검색되지 않았을 때)
R : 전체 관련영상 개수
K : 측정범위 $= min(4R,\ 2\times max\{R_j\})$
R_j: j번째 *data set*의 R

NMRR의 값을 얻기 위해서는 먼저 AVR을 계산해야 하며 그 다음 MRR을 계산하게 되는데 이는 평가값을 0부터 시작하게 하기 위함이며 최종적으로 NMRR이 얻어졌을 때는 [0,1]로 정규화 된다. 이 식은 한사람에 의해서 만들어진 것이 아니라 여러 사람에 의해 1년 가까운 시일에 걸쳐 수정 보완되어졌기 때문인지 단 한번에 정규화가 되지 않고 3단계에 걸쳐 번잡스러운 계산을 해야 한다.

6. 문제제기

지금까지 기존 평가척도들을 살펴보았는데 정확도-재현율 그래프는 그림 1처럼 상대비교가 곤란한 점이 있으며 복합척도도 몇 가지 문제점이 존재한다[1]. 또한 순위정보를 이용하는 척도가 아니므로 순위부여 시스템에 보다 적합한 평가방법이 필요하다. 표 1은 검색된 관련영상의 개수가 모두 같은 관계로 이들을 모두 같게 평가하는 RR과 PR의 문제를 보이고 있다.

표 1. RR과 PR의 문제
Table 1. Problems of RR and PR.

	1	2	3	4	5	6	7	8	9	10	RR	PR
a	×	×	×	×	×	×	×	○	○	○	3/3	3/10
b	×	×	×	○	×	○	×	×	×	○		
c	○	○	×	×	×	×	×	×	×	○		

BEP류 또한 RR-PR과 마찬가지 문제를 가지고 있다. 전체 관련영상이 3개라고 할 때 표 2의 유형 a, b, c 모두 검색되었지만 검색패턴은 제각기 다르다. 따라서 관련영상(○) 사이에 존재하는 부적합 영상(×)의 개수가 제각기 다르므로 이들을 구별할 수 있어야 하지만 측정범위에 따라 같게 평가되는 경우가 발생한다. 유형 b, c에서 각각 순위 7위, 4위 이후는 전체 관련영상 3개가 모두 검색된 이후이므로 이들(·)은 부적합 영상으로 보지 않는다.

표 2. BEP류의 문제
Table 2. Problems of a kind of BEP.

	1	2	3	4	5	6	7	8	9	···	20	FP	BEP	BEP_{1R}
a	×	×	×	○	×	×	×	×	○	×	○	33.3	33.3	0
b	×	×	×	○	○	○	·	·	·	·	·		100	
c	○	○	○	·	·	·	·	·	·	·	·	100		100

RR-PR과 BEP류와 달리 RP류는 순위값을 직접 이용함으로써 관련영상의 순위에 따른 분포 특성을 고려하고 있기는 하지만, 많은 문제점이 존재한다. 특히 AR의 경우 세 가지 문제점이 존재하는데 첫째는 정규화가 되어있지 않다는 점이며, 둘째는 검색된 관련영상의 개수(Rs)가 0일 때 계산불능이 되며, 셋째는 Rs가 다른 검색패턴들 간에는 측정결과가 부정확하다는 점이다. 이는 검색된 것만을 평가에 반영하기 때문인데 NAR 역시 AR의 세 번째 문제점 중 일부만을 해결할 뿐이다. 표 3에 전체 관련영상이 9개일 때 이러한 문제점을 나타내었는데 이러한 이유로 MR 또는 σ 를 함께 사용하기도 한다.

표 3. RP류의 문제
Table 3. Problems of a kind of RP.

	1	2	3	4	5	6	7	8	9	10	AP	NAR	MR	σ
a	○	○	○	○	○	○	○	○	○	·	5.0	1.0	9	2.582
b	○	×	×	×	○	×	×	×	○	×		2.5		3.266
c	×	×	×	×	○	×	×	×	×	×		5.0	5	0

표 4에 순위정보를 직접 이용하는 AR, NAR, NRS, NMRR의 공통 문제점이 나타나있는데, 전체 관련영상이 5개일 때이다. 세 유형에 대해 각 평가척도들은 모두 같은 평가값을 산출함으로써 이들을 구분하지 못한다. 즉, 한 관련영상이 +n 순위만큼 이동하면 다른 관련영상이 -n 순위만큼 이동하여 평균순위가 같은, 그러나 분포유형이 다른 검색패턴들은 구분할 수 없음을 의미한다.

표 4. AR, NAR, NRS, NMRR의 공통문제
Table 4. Common problems of AR, NAR, NRS and NMRR.

	1	2	3	4	5	6	7	8	9	10	AR	NAR	NRS	NMRR	#X
a	☆	×	○	○	△	×	◎	·	·	·					2
b	×	☆	○	○	◇	◎	·	·	·	·	4.0	1.33	0.75	0.125	1
c	◇	△	○	○	×	×	×	×	×	◎					5

지금까지 살펴본 기존 평가척도들의 문제점은 크게 세 가지로 설명할 수 있다. 하나는 측정인자로서 개수를 사용하는데 있다. 이는 측정범위 내 검색된 관련 또는 부적합 항목의 개수만 같으면 모두 같은 평가값을 산출하는 문제가 있다. RR-PR과 BEP류가 이에 해당하는데, 개수만을 측정인자로 이용한다는 것은 순위부여 시스템의 검색효율을 측정하기에는 부족하다. 따라서 개수가 아닌 순위를 직접 이용하는 AR, NAR, NRS, NMRR은 순위부여 시스템을 위한 평가척도라고 할 수 있다.

두 번째는 순위를 이용하는 AR, NAR, NRS, NMRR의 경우 표 4처럼 대칭패턴을 구별하지 못하는 경우이다. 이는 순위의 앞, 뒤를 균등하게 평가하기 때문에 발생하는 것으로 사용자의 관점에 비추어 앞쪽 순위에 더 큰 비중을 두어 평가해야 한다.

세 번째는 측정범위 설정 문제이다. BEP류가 개수를 사용하며 RR과 유사한데도 순위부여 시스템용으로 사용되는 것은 사용자의 입장에서 보았을 때 상위 앞부분이 중요하다는 사실에 따라 앞부분만 평가하기 때문이다. 즉, 질의결과가 매우 많을 경우 사용자는 앞쪽만 볼 확률이 높으므로 뒤쪽은 측정범위에서 제외하는 것이다. 그러나 이로 인한 새로운 문제를 표 5에 보이고 있다. BEP류와 NMRR은 실제 data set에 따라 측정범위가 결정되겠지만 표 5처럼 결정되었다고 가정하자. 유형 a는 관련영상을 6개 검색하였고 유형 b는 3개밖에 검색하지 못했으므로 유형 a가 더 우수하지만 측정범위 내에서는 결과가 달라진다. 이처럼 측정범위가 1순위 좁혀지느냐 혹은 넓혀지느냐에 따라서도 결과는 큰 차이를 보일 수 있다.

표 5. 측정범위 제한 문제
Table 5. Problem of limit to the range of measurement.

	범위 내					범위 밖				BEP류	NMRR	실제
a	…	○	○	×	○	○	○	○	…	a=b	b>a	a>b
b	…	○	○	○	×	×	×	×	…			

III. 제안한 단일가 평가척도

앞서 기존의 평가척도가 평가값 산출을 위해 개수를 이용하거나 측정범위를 제한하게 되면 그에 수반되는 많은 문제점들이 파생됨을 알 수 있었다. 이러한 문제를 해결하기 위해 본 장에서는 영상 데이터베이스 검색 알고리듬의 검색효율을 평가하기 위한 새로운 평가척도 NDS (Normalized Distance Sum)를 제안한다. 먼저 측정범위와 평가척도와의 상관관계를 살펴보고 검색효율이 좋은 시스템이 되기 위한 조건을 알아본다. 제안한 평가척도는 영상뿐만 아니라 순위가 부여되는 모든 미디어의 검색 시스템에 적용될 수 있지만 편의상 영상을 대표로 칭한다.

1. 측정범위에 대한 고려

평가값을 산출하기 위해서는 측정범위를 설정해야 하는데 그 범위에 따라 평가값이 달라질 수 있으며 많은 문제들이 내포되어 있다. AR류와 NRS는 전체범위를 평가대상으로 하지만 검색된 관련영상의 대상으로 평가하므로 측정범위가 마지막 관련영상의 순위를 넘더라도 평가값의 변화가 없다. 그러나 측정범위내 검색된 관련영상의 개수와 순위가 동일하더라도 범위값 변화에 따라 평가값이 변하는 척도로는 PR이외에 NMRR과 제안하는 NDS가 있으며 이들은 범위가 넓어질수록 상대적으로 평가가 좋아진다.

BEP류와 NMRR은 측정범위가 상위 일부분으로 제한되는데 이는 순위부여 시스템의 특성에 기인하는 것이지만 표 5의 문제가 발생할 수 있음을 뜻한다. 측정범위를 좁게하면 이와 같은 문제가 발생되는데, NMRR의 경우 측정범위 K밖의 관련영상에 대해서는 K+1 순위에 검색되거나 또는 검색되지 않거나 상관없이 모두 K+1 순위에 검색된 것으로 똑같이 간주하는 문제가 있다. MPEG-7 표준화 회의에서 약 1년여에 걸쳐 NMRR이 완성되어 가는 과정에서 측정범위 밖의 관련영상에 대한 순위를 K+1과 1.25K를 혼용하다가 최종 K+1로 채택하였으나 어느 쪽으로 하더라도 부정확한 문제로 남는다.

경계선을 어디로 할 것인가 하는 문제는 매우 민감한데 사용자의 입장에서 보면 일반적으로 자기가 보고자 하는 관련영상의 2배 범위를 넘지 않을 것이나 이 또한 고정된 것은 아니다. 그리고 MPEG-7 표준화 회의에서처럼 각 알고리듬의 우열을 가리기 위한 것이라면 각 검색패턴들의 검색된 관련영상의 대부분이 측정범위 내에 들어오도록 범위를 넓혀서 우열을 가리는데 정확성을 기해야 할

것이다.

MPEG-7의 Color/Texture 부문 표준 data set인 CCD(Common Color Data set)를 살펴보면 동영상의 각 프레임들이 포함된 경우가 있는데 이 경우 대다수의 알고리듬들에 의한 평가값은 매우 좋게 나타난다. 게다가 MPEG-7의 표준 평가척도인 BEP와 NMRR은 측정범위 내에 들어온 관련영상의 개수가 같기만 하면 평가값이 같거나 우열의 차이가 거의 나지 않는다. 반대로 CCD 일부는 검색이 어려워 뒤쪽에 검색되는 경우가 있는데 이 경우는 반대로 측정범위 밖에 나가게 되므로 우열을 가리는데 있어 부적절함을 알 수 있다. 이는 data set에 따라 측정범위 선정 기준이 달라져야 함을 뜻하므로 무조건 관련영상의 몇 배로 설정하는 것은 무리이다. 따라서 data set 설정시 data set 구성 내용과 알고리듬 기술수준에 따라, 관련영상의 n배로 설정하는 것이 합당할 것이다.

측정범위 설정시 너무 좁게 잡아 표 4,6의 문제를 일으키지 않도록 해야 하는 것은 물론 대부분의 관련영상이 포함되도록 충분히 넓혀야 하겠지만 하나의 data set 내에서도 그룹들 간 검색효율의 차이가 매우 심한 경우가 있을 수 있다. 즉, 검색패턴들의 마지막 관련영상의 순위 중 최소값과 최대값 차이가 크다면 상황에 맞추어 8할 정도만 포함시키는 것이 좋을 것이다.

다음으로 고려해야할 사항이 측정범위 내에서의 평가비중에 관한 것이다. 사용자가 순위부여 시스템의 질의결과를 보는 것은 뒤쪽 순위로 갈수록 보게 될 확률이 점차 낮아질 뿐 어느 시점에서 100%와 0% 확률의 경계선이 존재하는 것은 아니다. 그러나 기존 평가척도는 전체를 100%의 비중으로 평가하거나 상위 일부분의 경계선을 그어 범위 내에서는 100%, 범위 밖은 0% 비중으로 평가하는 이분법적인 처리방식을 따르고 있다. 그러나 과연 어디를 경계선으로 하는가 하는 문제가 발생한다. 좁게 잡으면 표 5의 문제가 발생하며 넓게 잡으면 순위부여 시스템의 특성을 따르지 않게 되어 표 4의 문제를 유발하므로 기존 평가척도로는 넓힐 수도 좁힐 수도 없는 딜레마에 빠진다. 따라서 제안하는 평가척도는 측정범위를 넓히고 순위에 따라 평가비중을 다르게 평가하도록 한다.

하나의 검색패턴이 아닌 여러 개의 검색패턴 결과를 상대 비교하기 위해서는 적어도 하나의 관련영상 그룹 내에서는 측정범위를 동일하게 설정해야한다. 다음은 일반적인 측정범위 설정에 관한 방법을 언급하고 있다.

[설정1] $R_{r_1}, R_{r_2}, \cdots, R_{r_n}$ ➡ 범위: $1 \sim R_{r_n}$ where R_{r_i}: i번째 관련영상의 순위

[설정2] $R_1, R_2, \cdots, R_n$ ➡ 범위: $1 \sim R_n$ where R_i : i번째 순위($R_n \le R_{r_n}$)

기존 척도들 중 측정범위의 끝이 [설정1]처럼 관련영상에 중심이 맞추어져 있는 것은 RP류, NRS로서 이들은 측정범위가 마지막 관련영상 Rr_n까지가 된다. [설정2]와 같은 경우는 BEP류, NMRR, 그리고 제안하는 척도 NDS이며 관련영상의 위치에 상관없이 임의의 순위로 고정된다.

NMRR과 NDS는 동일한 조건 하에서도 측정범위 값의 변화만으로도 평가값이 달라지므로 본 논문에서는 평가척도들을 비교하기 위해 제안할 평가척도의 측정범위를 통일 시켰다. 그러나 알고리듬의 비교가 목적이라면 NDS는 굳이 측정범위를 제한하여 표 5와 같은 문제를 일으킬 필요가 없다.

2. 설계지침

앞서 기존 평가척도들의 문제점과 측정범위에 대해서 살펴보았다. 측정범위와 평가비중과의 관계를 고려할 때 기존 평가척도로는 많은 문제점이 존재함을 알 수 있었다. 따라서 뒤쪽 순위로 갈수록 유사도가 낮은 영상들을 출력하는 순위부여 시스템의 특성을 따르기 위해 어떻게 평가해야 할 것인지는 자명해진다. 즉, 제안하는 평가척도의 설계지침은 측정범위를 넓히되 순위에 따라 평가비중을 가변적으로 감소시켜야 한다는 것이다.

또 다른 설계지침으로 평가척도가 지녀야할 평가조건은 무엇인가 하는 것이다. 평가척도는 검색효율을 정량화 하는 것이며 정보검색 시스템의 평가는 일반적으로 검색효율, 신속성, 경제성의 세 가지 측면에서 수행되는데 이 중 검색효율을 제외한 나머지 기준은 비교적 쉽게 측정할 수 있기 때문에 별 문제가 되지 않는다[1].

본 논문에서는 검색효율이 좋은 시스템이 되기 위한 검색조건을 다음과 같이 정의하였는데 [검색조건3]~[검색조건4]는 순위부여 시스템을 위해 추가적으로 적용되는 사항이다. 좋은 검색 시스템이란 다음의 [검색조건]들을 가급적 최대한 만족하는 시스템을 말하며 이를 측정하기 위한 평가척도 역시 다음의 [평가조건]들을 모두 만족해야 한다.

[검색조건1] 많은 관련영상을 검색할 수 있어야 한다.(RR이 커야 한다.)
[검색조건2] 검색된 부적합 영상이 적어야 한다.(PR이 커야 한다.)
[검색조건3] 관련영상이 앞쪽 순위에 있어야 한다.(RP류는 작고 BEP류는 커야 한다.)
[검색조건4] 관련영상이 몰려있어야 한다.(σ 와 MR이 작아야 한다.)

[평가조건1] [검색조건1~4]를 모두 평가할 수 있어야 한다.
[평가조건2] 하나의 척도로 모든 [검색조건]을 평가하는 단일가 척도이어야 한다.
[평가조건3] 0~1사이의 정규화된 값이어야 한다.
[평가조건4] 앞쪽에서 뒤쪽으로 갈수록 비중을 낮추어 평가해야 한다.

초기 검색 알고리듬은 [검색조건1]의 수치를 높이는데 주력하지만 일정 기간이 지나면 이보다는 [검색조건2]에 더 큰 의의를 두게 된다. 이용자의 입장에서도 검색된 영상이 수백 수천 개가 될 때 부적합한 영상은 가급적 덜 보고 싶어할 것이다. 또한 순위부여 시스템에서는 유사도 크기에 따라 출력하므로 실제 관련영상들은 가급적 앞쪽 순위에 출력될수록 이상적인 시스템이 된다.

[검색조건1]이 같은 상황에서 [검색조건4]만 놓고 보았을 때는 주관적이라고 생각할 수도 있으나 이는 [검색조건2]와 [검색조건3]보다 우위에 있지 않다. 즉, 표 1과 같은 경우 유형 'a'가 몰려있으나 [검색조건3]에 의해 나쁘게 평가된다. 표 4의 경우는 유형 'b'가 몰려있으며 [검색조건2]를 충족하므로 좋게 평가된다. 이는 평균순위는 같더라도 관련영상이 몰려 있어야 부적합 영상이 적어 사용자가 보는 인지적 노력이 감소하기 때문이다.

기존의 평가척도들과 달리 제안하는 방법은 앞서 언급한 조건들을 모두 평가에 반영할 수 있도록 설계하였다. 이는 같은 측정범위 내에서 관련영상의 개수가 같은 경우라도 검색패턴에 따라 보다 세밀하고 정확하게 측정할 수 있어야함을 뜻한다. 따라서 기존척도들과 다른 점은 [검색조건1]보다는 [검색조건2]에 더 큰 비중을 두며 [검색조건4]와 [평가조건4]가 새로이 내포된 단일가 척도라는 점이다.

3. 제안한 평가 척도(NDS)

순위부여 시스템에서는 순위가 낮아질수록 검색된 영상을 보아야 하는 노력도 비례해서 증가하게 되는데, 임의의 관련영상이 위치한 순위 m까지 보는데 필요로 하는 노력은 m개의 관련 혹은 부적합 영상 하나 하나를 보는데 드는 노력의 합이 된다. 따라서 우수한 시스템이란 순위 m까지의 영상을 보는데 드는 노력을 최소화할 수 있어야 한다.

m개의 영상에는 부적합 영상도 함께 포함되어 있으므로 이들의 개수를 최소한으로 줄일 필요가 있다. 또한 부적합 영상의 개수가 최소화된 상태라 할지라도 부적합 영상의 위치에 따라 노력의 크기는 달라진다. 즉, 부적합 영상이 높은 순위에 올수록 관련영상은 낮은 순위에 오게되며 이는 부적합 영상으로 인해 관련영상을 보는 노력이 증가함을 뜻한다.

관련영상을 보는 노력은 당연하지만 부적합 영상의 개수와 그 위치로 인해 관련영상을 보는 노력이 커질 수 있으므로 이러한 불필요한 노력을 최소화할 수 있어야 한다. 따라서 제안하는 방법은 시스템이 검색한 모든 관련영상 사이에 분포하고 있는 부적합 영상들의 개수뿐만 아니라 그 위치를 성능평가의 기준으로 삼아 앞서의 불필요한 노력을 보다 정확하게 정량화한 것이다.

측정범위가 m순위까지이고 부적합 영상으로 인한 불필요한 노력을 "부적합 거리"라 칭할 때 이에 대한 예는 표 6과 같다. 이는 순위에 반비례하는 값으로 임의의 부적합 영상의 부적합 거리 d_i는 식 (7)로 정의하며, 측정범위 1~m 순위를 0~1로 정규화한 r_{ni} 및 이에 따라 정규화된 거리함수 d_{ni}를 식 (8)과 같이 정의한다.

표 6. 순위에 따른 거리
Table 6. Distance according to ranked position.

순위	1	2	3	···	i	···	m-2	m-1	m	···
d_i	m	m-1	m-2	···	m-i+1	···	3	2	1	0

$$d_i = m - i + 1 \tag{7}$$

$$r_{ni} = (i-1)/m \qquad d_{ni} = 1 - r_{ni}$$

$$where \quad 1 \le i \le m \rightarrow 0 \le r_{ni} < 1 \quad \begin{cases} i = 1 \rightarrow d_{ni} = 1 \\ 2 \le i \le m \rightarrow 0 < d_{ni} < 1 \\ i > m \rightarrow d_{ni} = 0 \end{cases} \tag{8}$$

m 순위의 영상을 보기 위한 노력 중 부적합 영상들로 인한 불필요한 노력은 관련영상들 사이에 존재하는 부적합 영상들의 부적합 거리를 누적하여 정규화하면 된다. 이때 부적합 거리는 순위를 인자로 하는 다양한 거리함수를 사용하여 나타낼 수 있는데 식 (9)와 그림 1에 그 일 예

$$d_{ni} = \begin{cases} f_i(r_{ni}) = 1 - r_{ni} & r\text{의 역, 가장 일반적} \\ f_0(r_{ni}) = \sqrt{1 - r_{ni}^2} & (0,0)\text{원, 뒤쪽에 민감} \\ f_1(r_{ni}) = -\sqrt{1 - (r_{ni}-1)^2} + 1 & (1,1)\text{원, 앞쪽에 민감} \\ f_c(r_{ni}) = (1 + \cos \pi r_{ni}) / 2 & \cos r, \text{ 가운데 민감} \end{cases} \tag{9}$$

$$where \quad 1 \le i \le m \rightarrow 0 \le r_{ni} < 1 \rightarrow 0 < d_{ni} \le 1$$

를 나타내었다. 여기서 $f_i(r_{ni})$를 제외한 나머지 거리함수들은 일반적인 경우가 아니지만 밀집도 같은 어떤 특성에 민감하게 반응하도록 쉽게 확장될 수 있다.

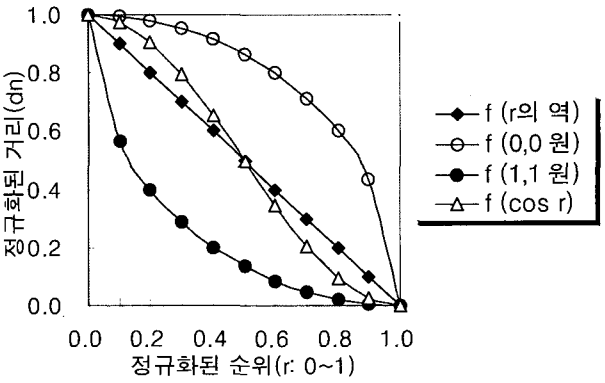

그림 1. 거리함수
Fig. 1 Distance function.

이제 NDS는 식 (10)과 같이 정의된다. 이는 단 하나의 평가항목으로 기존 평가척도들의 문제점을 해결할 수 있는 새로운 단일가 척도이다. 0이상 1이하의 값을 가지며 작을수록 좋은 검색효율을 나타낸다. 0인 경우는 모든 관련영상이 1순위부터 차례로 검색된 경우이며 1인 경우는 관련영상이 하나도 검색되지 않은 경우이다.

$$NDS(m) = \frac{\sum_{i=1}^{l}(m-i+1 \mid i\text{:부적합 영상의 순위})}{m(m+1)/2} \quad \text{or}$$

$$= \frac{\sum_{i=1}^{l}(d_{ni} \mid i\text{:부적합 영상의 순위})}{\sum_{i=1}^{m} d_{ni}} \qquad (10)$$

$$where\ l = \begin{cases} r_L & \text{if } r_L < m \\ m & \text{else} \end{cases}, \quad 1 \le l \le m$$

r_L: 마지막 관련영상의 순위
m: 측정범위의 마지막 순위
= 전체 관련영상의 n배(nR)
$d_{ni} = f(r_{ni})$: 원하는 거리함수

표 7은 많이 사용되는 기존 평가 척도들과 제안한 척도를 비교하기 위해, 측정범위 m=9, DB내 영상의 개수가 20, 전체 관련영상의 개수 R=3, RR=1일 때 이의 분포유형에 따른 평가값을 나타낸 것이다. 표 7안의 음영 영역의 수치는 평가가 제대로 이루어지지 않았음을 뜻하는 것이다. 제안한 식 (10)으로부터의 평가척도 NDS를 제외하고는 최소한 한번 이상은 서로 다른 유형임에도 불구하고 같은 평가값 산출한다. 또한 유형 'a'와 'd'의 경우 AR, NAR은 유형 'd'가 더 우수한 것으로 평가하지만 NDS는 PR, MR, σ 의 특성이 반영되어 있기 때문에 유형 'a'를 더 우수한 것으로 평가한다.

제안한 평가척도인 NDS의 장점은 다음과 같다. 특히 기존의 평가척도들이 앞서의 [평가조건]들을 모두 고려하지 않아 발생할 수 있는 여러 문제점들을 NDS는 거리함수를 사용함으로써 해결한다는 점이 기존 방법과 차별되는 요소이다.

[장점1] [평가조건1~4]를 모두 고려(특히 [검색조건4], [평가조건4])
[장점2] 세밀한 측정이 가능(RR-PR, BEP류는 개략적, ∵개수)
[장점3] 대칭적인 검색패턴도 구별함
[장점4] 평가DB의 특성에 따라 다양한 거리함수 선택 가능

지금까지 기존 여러 평가척도들과 제안한 NDS에 대해 살펴보았다. 평가척도에 사용되는 인자(argument)를 기준으로 할 때 NDS가 기존 평가척도들과 다른 점은 BEP류처럼 측정범위 제한에 의한 순위의 간접정보도 아니고 AR류, NRS, NMRR처럼 순위정보를 직접 이용하는 것도 아닌 순위에 따라 반응하는 거리함수를 이용한다는 점이다. 이로서 평가척도는 측정인자와 측정범위를 기준으로 다음과 같이 변화되어 왔음을 알 수 있다.

[측정인자] 개수 ➡ 순위 ➡ 거리
[측정범위] 전체 ➡ 앞쪽 ➡ 전체
(비중) (균일) (균일) (가변)

표 7. RR=1일 때 각 평가척도에 의한 평가
Table 7. Evaluation according to each measure at RR=1.

유형	관련영상의 분포 1 2 3 4 5 6 7 8 9	PR	FP	BEP	BEP_{IR}	AR	NAR	MR	σ	NRS	NMRR	NDS(순위)
a	× × × ○ ○ ○ · · ·	3/6	1/3	100	0	5.00	2.50	6	0.82	0.40	0.38	0.533(3)
b	○ × × ○ × ○ · · ·				33.3	3.67	1.83		2.06	0.55	0.21	0.444(1)
c	○ × ○ × × × ○ · ·	3/7	2/3	66.7	66.7			7	2.49			0.511(2)
d	○ × ○ × × × × × ○	3/9				4.33	2.17	9	3.40	0.46	0.29	0.622(4)
e	× × × × × × ○ ○ ○		0/3	0	0	8.00	4.00		0.82	0.25	0.75	0.867(5)

IV. 실험결과

제안한 NDS와 기존 평가척도와의 비교실험에는 두 가지 목적이 있다. 하나는 기존 평가척도들이 갖고 있는 문제점들을 검색패턴에서 찾아내는 것이고 또 다른 하나는 평가척도들의 평가값 분포곡선을 통해 그 특성을 파악하는 것이다. 이를 위해서는 가급적 모든 검색패턴을 대상으로 실험해야 하지만 실제 데이터베이스를 검색한 하나의 검색패턴 집합으로는 극히 일부분의 검색패턴만을 얻을 수 있을 뿐이다. 이로부터는 기존 평가척도들의 오류를 골고루 발견하기 힘들뿐만 아니라 각 평가척도들의 평가값 분포곡선 또한 얻을 수 없다. 따라서 본 실험에서는 주어진 조건하에서 발생 가능한 모든 검색패턴을 자동생성 시키는 패턴생성기를 만들고 이를 대상으로 시뮬레이션 결과를 보인다.

1. 패턴생성기

n순위 내에서 r개의 관련영상이 위치하는 모든 경우의 수는 조합(combination)의 수 ${}_nC_r(= n! / r!(n-r)!)$로 나타낼 수 있다. ${}_nC_r$개의 패턴을 생성시키기 위해서 먼저 각 패턴의 발생순서를 식 (11)과 같이 정의한다. 패턴은 s_1부터 s_m까지 ${}_nC_r$개가 차례로 생성되며 이들 집합을 L_s로 표현한다. 하나의 패턴 s_i는 L로 표현되는데 이는 '•'와 'O'으로 표기되고 'O'가 r개 발생하며 'O' 앞뒤로 '•'이 위치할 수도 있음을 뜻한다. 여기서 '•'은 부적합 영상[1]을, 'O'은 관련영상을 뜻하는 것으로 s_i는 이러한 개별 구성요소 a_{ij}가 n개로 구성된 것이다. 각 패턴의 발생순서는 a_{ij}에 따라 결정되는데 'O'가 '•'보다 먼저 오며 s_2가 s_1보다 뒤에 발생되는 것을 $s_1 < s_2$로 표현한다면 각 패턴의 발생 순서 규칙은 식 (11)과 같이 표현된다.

$$
\begin{aligned}
&L = (\cdot^* \bigcirc \cdot^*)^r \\
&L_s = \{s \mid s \in L \text{ and } |s| = n\} \\
&\text{For } s_1, s_2 \in L_s, \text{ Let } s_1 = a_{11}a_{12}\cdots a_{1n} \text{ and } s_2 = a_{21}a_{22}\cdots a_{2n} \\
&\text{For } a_{11} = '\bigcirc',\ a_{21} = '\cdot',\ \text{Let } a_{11} < a_{21} \\
&s_1 < s_2 \text{ If } (a_{11} < a_{21}) \text{ or } (a_{1j} = a_{2j} \text{ for } j = 1, 2, \cdots, k-1 \text{ and } a_{1k} < a_{2k})
\end{aligned}
\tag{11}
$$

1) 관련영상이 모두 위치한 이후엔 부적합 영상이 아니지만 패턴발생시에는 이들을 구별 안함

그림 2는 식 (11)의 규칙을 따르는 패턴생성 알고리듬으로 재귀적 호출(recursive call)에 의해 구현한 것이다. 이는 주어진 조건하에서 발생 가능한 모든 패턴을 자동 발생시키며, 실제 프로그램에서는 그 패턴들을 대상으로 각 평가척도들에 의한 평가값들도 자동 계산되도록 하였다.

```
void main()
{
   GetArgument(&n, &r);
   combi(0, n, r);
}

void combi(int s, int n, int r)
{
if (r==1) {
   for (i=s+1; i<=s+n; i++) {
      rank[i] = 'O';
      //패턴출력 후 마지막 O의 순위 기억
      rlr = DispPattern();
      CalcEval(rlr); //평가값 계산
      rank[i] = '•';
   }
}else{
   for (i=1; n-i>=r-1; i++) {
      rank[s+i] = 'O';
      combi(s+i, n-i, r-1);
      rank[s+i] = '•';
   }
}
}
```

그림 2. ${}_nC_r$ 패턴생성 알고리듬
Fig. 2. ${}_nC_r$ pattern generation algorithm.

2. 검색패턴의 조건설정

${}_nC_r$개의 패턴을 발생시키기에 앞서 조건 설정에 2가지 기준을 적용하였다. 첫째는 발생시키고자 하는 패턴의 개수에 대한 것이고 둘째는 발생된 측정범위에 대한 것이다. 먼저 ${}_{10}C_5(=252)$개는 측정오류를 발견하기 위해, 그리고 ${}_{20}C_9(=167{,}960)$개는 분포유형을 알아보기 위해 각각 발생시켰다. 다음으로 측정범위 설정은 전체범위와 임의의 범위로 나누어 측정하였다. 전체범위는 전체 관련영상이 모두 검색되어 나타난 패턴을 뜻하며 임의의 범위로 축소하면 검색된 관련영상의 개수가 서로 다른 패턴을 대상으로 측정함을 의미한다.

범위 축소시 그 범위가 BEP류와 NMRR의 측정범위라고 가정한다면 이들은 범위 밖을 고려하지 않으므로 표 5

의 문제가 발생한다. 더구나 BEP류와 NMRR의 측정값 그 자체는 전체 측정범위를 기준으로 보면 모두 부정확한 값이 된다. 따라서 이 문제는 여기서 더 이상 언급하지 않는다. 실험에서는 검색된 관련영상의 개수가 서로 다른 경우를 대상으로 패턴들간의 우열문제와 평가값 분포를 알아보기 위해 범위를 축소하고 각 평가척도들의 측정범위를 축소된 범위로 일치시켜 측정한다.

3. 측정오류 검출

기존 평가척도의 오류를 검출하기 위하여 ${}_{10}C_5(=252)$개의 패턴을 발생시켰는데 측정범위와 관련영상의 개수가 2:1이 되도록 설정한 이유는 모든 검색패턴에 대해 BEP는 100%를 나타내어 구별할 수 없음을 보이기 위해서이다.

먼저 측정범위를 전체로 하며 252개 모든 패턴에 대해 RR=1.0, BEP=100이 되므로 이들 척도로는 구별할 수 없는 패턴이 된다. 그래서 순위를 이용하는 평가척도들 중 대표로 AR을 선택하여 NDS와의 평가값을 비교하여 그림 3을 얻었다.

그림 3에서 수직선들은 NDS의 평가값을 고정하였을 때 이에 대한 상대 평가척도 AR의 평가값 변화폭을 나타낸 것으로 수직선 하단은 최소값을 갖는 패턴이, 상단은 최대값을 갖는 패턴이 오며, 사각점은 상대 평가값의 산술평균을 나타낸다.

AR과 NDS가 별 차이가 없다면 x축 증가에 따라 선형적으로 y축 값도 증가해야 하지만 그림 2에 나타난 수직선들로 인해 서로간의 차이가 존재함을 알 수 있다. 즉, 임의의 x값에 해당하는 수직선 최대값 y_m에 대해 그 다음 x값의 수직선에 존재하는 모든 값은 y_m보다 커야하지만 그렇지 않은 경우가 많다. 다시 말해 y_m을 원점으로 하였을 때 4/4분면에 해당되는 패턴들은 y_m에 해당되는 패턴과 서로 상반되는 결과로, 인지적으로 우열을 가리기 힘든 경우를 제외하고는 둘 중 하나의 평가척도가 부정확한 경우라고 말할 수 있다.

그림 3에서 AR의 4.0에 해당하는 각 패턴 A~E와 F에 대한 평가값들을 표 8에 나타내었다. 패턴 A~E에 대해 AR을 비롯 NAR, NRS, NMRR은 절대값만 다를 뿐 이들을 구별하지 못하지만 NDS는 A를 가장 우수한 것으로 평가한다. 이는 표 4에 해당하는 경우로 관련영상 5개를 사용자가 보기 위해서 패턴 A는 순위 6위까지 보아야 하며 이때 부적합 영상 하나를 불필요하게 보아야 하지만

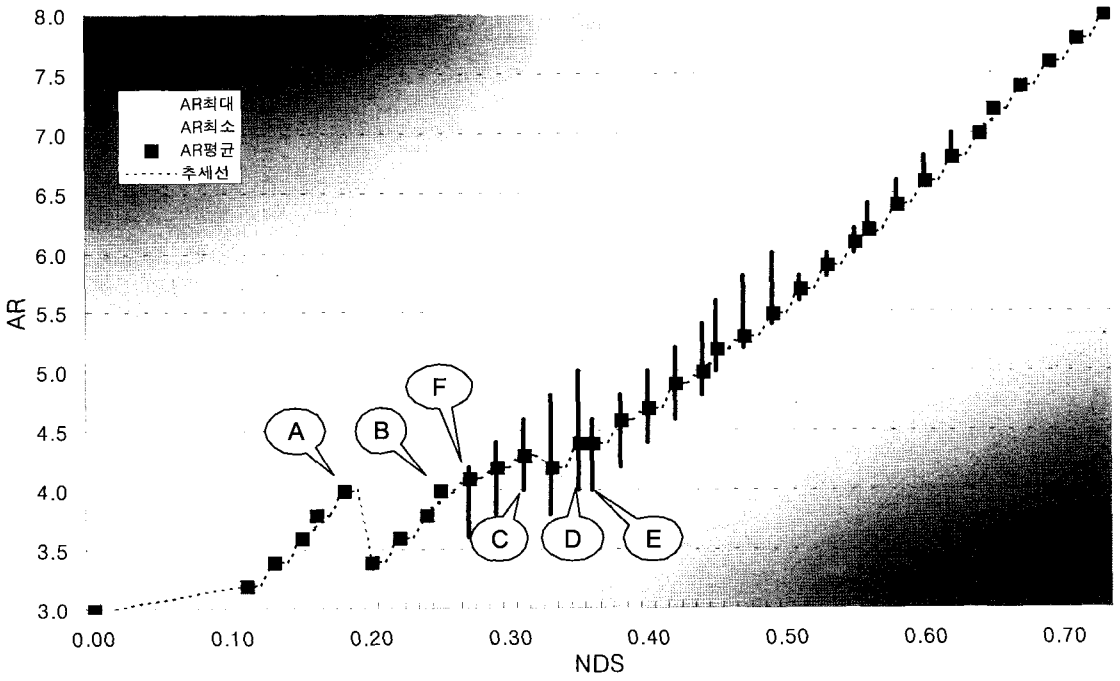

그림 3. ${}_{10}C_5$에서 AR과 NDS와의 관계
Fig. 3. Relation between AR and NDS at ${}_{10}C_5$.

표 8. 그림 3의 패턴 A~F에 대한 AR, NRS, NMRR, NDS, PR
Table 8. AR, NRS, NMRR, NDS, and PR of the pattern A~F in Fig. 3.

유형	패 턴	AR	NAR	NRS	NMRR	NDS	PR
A	×○○○○○·····	4.0	1.33	0.75	0.13	0.182	0.83
B	○×○○○×○····	4.0	1.33	0.75	0.13	0.255	0.71
C	○○×○○××○··	4.0	1.33	0.75	0.13	0.309	0.63
D	○○○×○××××○·	4.0	1.33	0.75	0.13	0.346	0.56
E	○○○○××××××○	4.0	1.33	0.75	0.13	0.364	0.50
F	○○××○○○····	4.2	1.40	0.71	0.15	0.273	0.71

패턴 E는 순위 10위까지 보면서 부적합 영상 5개를 보아야 한다. 또한 NDS는 유형 E보다 F를 더 좋은 것으로 평가하였지만 순위를 이용하는 척도들은 반대로 평가하였다. 표 8에 참고적으로 PR을 함께 나타내었는데 이로부터 알 수 있듯이 NDS는 PR의 성격을 강하게 내포하고 있음을 알 수 있다.

다음으로 측정범위를 상위 일부분으로 축소함으로써 검색된 관련영상의 개수가 다른 패턴들을 대상으로 측정하였는데 이 때 측정범위 K내에 검색되지 못한 관련영상의 순위에 대해 NRS의 경우 NMRR과 같이 K+1로 설정하였다. 반면 BEP류, AR, NAR은 검색된 영상만을 대상으로 하며, NDS는 검색된 부적합 영상을 대상으로 하므로 해당사항이 아니다.

측정범위는 절반인 5위까지(r5)로 축소하였는데 이 경우 중복되는 패턴을 제외하면 전체 252개 패턴에서 32개로 되며 이들에 대해 순차적으로 번호를 부여하였다. 이때 BEP_{1R}(100을 곱하지 않은 경우임)과 RR, PR이 모두 같게 되며 NMRR과 NDS 또한 같게 된다.

기존 평가척도들의 문제점을 예시하기 위해 32개 패턴을 조사해본 결과 매우 많은 문제점들이 나타났는데 표 9, 10은 그 중 일부만을 예시한 것이다. BEP_{1R}/RR/PR의 경우 패턴 1, 32를 제외한 나머지에 대해 평가값이 0.8(패턴 5개), 0.6(10개), 0.4(10개), 0.2(5개)로 세밀하지 못한 평가결과가 나왔다. 또한 AR은 NAR의 모든 오류를 포함하며 그 외의 오류도 많으므로 AR만의 오류는 따로 예시하지 않고 표 10에 AR과 NAR 공통 오류 일부를 예시하였다. 즉, 패턴 1, 2, 4, 8, 16의 경우 AR은 반대로, NAR은 모두 같게 평가하였으며 패턴 16, 3과 패턴 6, 5의 경우 AR, NAR 모두 반대로 평가하는 오류가 있었다. 게다가 패턴 32는 AR, NAR의 분모가 0이 됨으로 인해 계산불능이었다.

4. 평가값 분포특성

지금까지 살펴본 바에 의하면 순위부여 시스템의 검색효율을 평가하기 위한 척도들 중에서 AR류는 매우 많은 문제점들을 드러내었으며 BEP류 역시 매우 개략적으로 측정하는 문제가 있었다. 이제 평가값 분포특성을 알아보기 위해 $_{20}C_9$(=167,960)개의 패턴에 대해 측정하되 전체 범위(20위)에 대한 BEP값은 모두 100%이므로 비교적 세밀하게 측정하는 NRS와 NMRR, 그리고 제안한 NDS를 대상으로 조사하였다. 그러나 167,960개의 평가값을 차트로 나타낼 수 없어서 매 100번째마다 발췌하여 1,679개 패턴에 대해 그림 4로 나타내었다.

표 9. BEP1R/RR/PR의 문제
Table 9. Problems of BEP1R/RR/PR.

No.	pattern	BEP_{1R}/RR/PR	NDS
2	○○○○×	0.8	0.067
17	×○○○○	0.8	0.333
4	○○○××	0.6	0.200
25	××○○○	0.6	0.600
8	○○××××	0.4	0.400
29	×××○○	0.4	0.800
16	○××××	0.2	0.667
31	××××○	0.2	0.933

표 10. AR과 NAR의 문제
Table 10. Problems of AR and NAR.

No.	pattern	AR	NAR	NDS
32	×××××	불능	불능	1.000
1	○○○○○	3.00	1.00	0.000
2	○○○○×	2.50	1.00	0.067
4	○○○××	2.00	1.00	0.200
8	○○××××	1.50	1.00	0.400
16	○××××	1.00	1.00	0.667
3	○○○×○	2.75	1.10	0.133
6	○○×○×	2.33	1.17	0.267
5	○○×○○	3.00	1.20	0.200

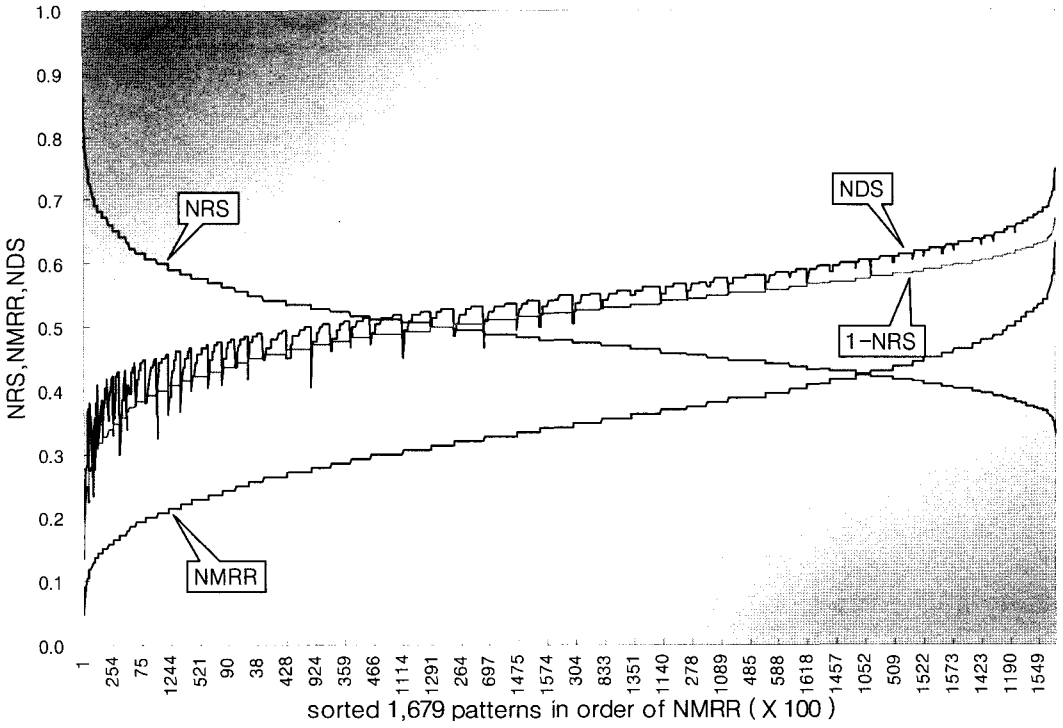

그림 4. $_{20}C_9$개 패턴에서 매 100번째 패턴들에 대한 NRS, NMRR, NDS의 평가결과
Fig. 4. Evaluation result of each 100th patterns at 20C9 by NRS, NMRR, and NDS.

표 8에서 보았듯이 관련영상의 개수가 같고 측정범위가 같은 경우 AR, NAR, NRS, NMRR은 유사한 특성을 보였는데 그림 4에서 보듯이 이들의 분포곡선의 변화점은 똑같다. 따라서 NDS와 이들은 그림 4에서처럼 차이를 보이며 이는 표 8의 문제점을 의미한다. 또한 앞쪽 순위에서는 NDS가 기존척도와 차이를 크게 보이며 뒤로 갈수록 차이가 줄어드는 것은 [평가조건4]를 따르기 때문이다.

NDS와 NMRR이 0부터 시작하지 않고 NRS도 1부터 시작하지 않는 이유는 첫 번째 평가값이 167,960개 패턴에서 100번째 위치한 패턴부터 시작되었기 때문이며 뒤쪽의 경우 또한 같은 이유에다가 관련영상이 8,7,…,1,0개인 경우의 패턴들이 제외되었기 때문이다.

평가값 변화시점은 다르지만 NRS와 NDS는 평가값 분포가 유사한 반면 NMRR은 상대적으로 낮은 수치를 보이고 있다. 이는 관련영상이 모두 검색되기만 하면 부적합 영상이 얼마나 포함되어 있던지 간에 대체로 좋게 평가함을 의미한다.

다음으로 그림 4의 1,679개 패턴에 대해 측정범위를 9위까지(r9)로 제한하였는데 이때 중복된 패턴을 제거하고 남은 908개의 패턴에 대한 측정한 결과를 그림 5에 보이고 있다. 이 패턴들은 관련영상이 0개에서 9개 전부 검색된 경우까지 다양하게 존재하므로 평가값은 0에서 1사이에 고르게 분포해야 하지만 NRS의 경우 한쪽으로 치우쳐져 있음을 알 수 있다. 그림 5는 평가값 분포를 보기 위해 측정범위를 똑같이 축소한 상태에서 측정한 것이며 실제 각 척도의 측정범위는 각각 다르게 설정되어 그림 5와는 다르게 된다.

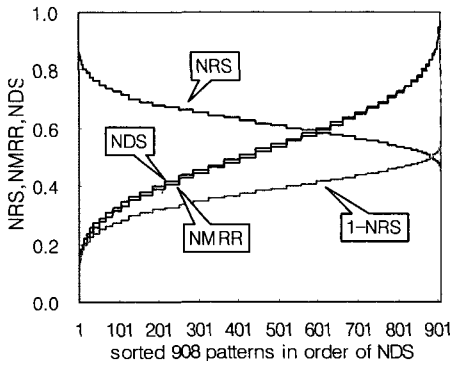

그림 5. 그림 4의 r9에서 NRS, NMRR과 NDS와의 관계
Fig. 5. Relation among NRS, NMRR, and NDS at r9 of Fig. 4.

V. 결 론

본 논문에서는 순위가 부여되는 영상 데이터베이스 검색 시스템의 검색효율을 평가하기 위한 새로운 단일가 평가척도 NDS를 제안하였다. NDS는 측정범위 내 부적합 영상의 순위에 따른 거리에 기반을 둔 평가방법으로 재현율과 정확도, 앞쪽 순위로의 군집성 등과 같은 특성이 모두 고려되었을 뿐만 아니라 평가비중도 기존의 이분법적이 아닌 순위에 따라 점차 낮추어 가는 평가방법이다. 또한 재현율보다는 정확도에 점차 더 큰 비중을 두는 요즘 추세에 적합하다고 할 수 있다.

기존 평가척도들은 이러한 여러 특성 중 일부만을 반영하고 있으며 특히 AR류는 매우 부정확한 결과를 보였다. MPEG-7 공식 평가척도 중 하나인 BEP는 평가결과가 매우 개략적인 것으로 나타났으며 이와 함께 NMRR은 측정범위를 제한하는데서 오는 여러 문제로 인하여 평가척도가 지녀야 할 특성을 모두 반영하고 있지 못했다. NRS는 관련영상이 모두 검색되지 않은 경우 정규화에 문제가 있음을 알 수 있었다.

이에 반해 NDS는 측정범위를 제한하지 않고 순위에 따라 비중을 낮추어 평가함으로써 기존 평가척도들의 문제점들을 모두 해결하였으며, 평가방법 또한 정확하고 간단하다. 임의의 검색패턴을 자동 발생시켜 측정해 본 결과 기존 척도의 문제점들과 NDS의 장점들을 예시할 수 있었으며 이는 제안한 방법의 우수성을 입증하는 것이다.

참 고 문 헌

[1] 정영미, 정보검색론, 구미무역, 1993.2

[2] N. Jardine and C.J. van Rijsbergen, "The Use of Hierarchic Clustering in Information Retrieval," *Information Storage and Retrieval 7*, no. 5, pp. 217-240, 1971.

[3] D. Billsus and M. J. Pazzani, "Learning Collaborative Information Filters," *Proceeding of Workshop on Recommender Systems*, 1998.

[4] W. S. Cooper, "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems," *American Documentation 19*, no. 1, pp. 30-41, 1968.

[5] Asha Vellaikal and C.-C. Jay Kuo, "Content-Based Image Retrieval Using Multiresolution Histogram Representation," *SPIE Digital Image Storage and Archiving Systems*, pp. 312-323, Philadelphia, Oct. 1995.

[6] S. Jeannin and M. Bober, "Description of Core Experiments for MPEG-7 Motion/Shape," *ISO/IEC JTC1/SC29/WG11 MPEG99/N2690, Seoul*, Mar. 1999.

[7] 최정식, 유광석, 김회율, "경계선 기반 색상 Spatiogram을 이용한 내용기반 영상 검색", *한국통신학회 하계종합학술발표논문*, vol. 17, no. 1, pp. 382-385, 1998.7

[8] Markus Stricker and Alexander Dimai, "Spectral Covariance and Fuzzy Regions for Image Indexing," Machine Vision and Applications, vol. 10, pp. 66-78, 1997.

[9] Ju Han and Kai-Kuang Ma, "A Novel Color Histogram Representation for Color Images," *ISO/IEC JTC1/SC29/WG11 M5510, Maui*, Dec. 1999.

[10] V. V. Vinod, "Description of Core Experiments for MPEG-7 Color/Texture," *ISO/IEC JTC1/SC29/WG11 MPEG99/N2929*, Melbourne, Oct. 1999.

저 자 소 개

서 창 덕

1998년 2월 : 한양대학교 전자공학과 학사
1999년 8월 : 한양대학교 전자공학과 석사
2000년 8월 : 한양대학교 전자공학과 박사 예정
주관심분야 : 멀티미디어 정보검색

김 회 율

1980년 : 한양대학교 전자공학과 학사
1983년 : Pennsylvania State University 전기공학과 석사
1989년 : Purdue University 전기공학과 박사
1989년 9월 ~ 1994년 2월 : University of Texas 조교수
1994년 현재 : 한양대학교 부교수로 재직
주관심분야 : 영상처리, 컴퓨터비전, 패턴인식