Diversity Conscious Retrieval

David McSherry

from

“Advances in Case Based Reasoning”,
proceedings ECCBR-2002,
Eds: Susan Craw and Alun Preece

Slides by U.Senthil, AIDB Lab

Why Diversity?

- Cases that are similar to the target query also tend to be very similar to each other
- Need for recommender systems to offer a more diverse choice of alternatives
- Need for more diversity conscious approach is highlighted by the growing trend towards the use of internet-enabled mobile phones, with screen size capable of displaying only a few recommendations

Standard Retrieval Set (SRS)

- The set of cases that are retrieved and presented as alternatives to the user is known as the *retrieval set*

- In a typical recommender system, the standard retrieval set (SRS) for a target query consists of the *k* cases that are that are most similar to the target query

Measures of Similarity and Diversity

$$similarity\,(R) = \frac{\sum_{i=1}^{k} Sim(C_i, Q)}{k}$$

$$Sim_{MF}(C,Q) = \frac{\left|\left\{a \in A_Q : \prod_a(C) = \prod_a(Q)\right\}\right|}{\left|A_Q\right|}$$

Matching Features

Note : Boolean Unweighted Match

$$diversity\,(R) = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\left(1 - Sim(C_i, C_j)\right)}{k\,\frac{(k-1)}{2}}$$

Increasing Diversity

- The process of constructing a retrieval set for a given query that is more diverse than the SRS for that query is know as *diversification.*

- Select next case *C* such that its relative diversity *w.r.t. R = {C*$_1$, *C*$_2$, … *C*$_n$*)* is the highest.

$$relative\_diversity(C,R) = \frac{\sum_{i=1}^{n}\left(1 - Sim(C, C_i)\right)}{n}$$

Example Case Library

Query

Selected

Case No.	beds = 4	Style = det	Rec	Loc = A	Sim_{MF}	SRS	BG	DCR-1
29	4	det	2	A	1.00	1	1	1
5	4	det	2	A	1.00	1	0	1
48	2	det	2	A	0.67	1	1	1
40	3	det	2	A	0.67	1	1	0
38	3	det	1	A	0.67	1	0	0
31	4	sem	2	A	0.67	0	0	1
16	4	det	1	C	0.67	0	1	1
8	4	sem	2	A	0.67	0	0	0
50	4	sem	2	D	0.33	0	1	0
49	2	det	2	C	0.33	0	0	0

Similarity:	0.80	0.67	0.80
Diversity:	0.26	0.60	0.40

Example (contd..)

$$similarity\left(BG\right)=\frac{1.00+0.67+0.67+0.67+0.33}{5}=0.67$$

$$diversity\left(BG\right)=\frac{\left(0.33+0.33+0.33+0.67\right)+\left(0.33+0.67+1\right)+\left(0.67+1\right)+\left(0.67\right)}{5\times\frac{\left(5-1\right)}{2}}=0.60$$

BG has increased diversity from 0.26 to 0.60. However, the increase in diversity (0.34) is much greater than the corresponding loss in similarity (0.13).

Diversity 1 – Bounded Random Selection

```
t: target query, C: case-base, k: # results, b: bound
define BoundedRandomSelection (t, C, k, b)
begin
    $C'$=bk cases in C that are most similar to t
    R=k random cases from $C'$
    return R
end
```

Retrieve b times more cases
Choose k from them randomly

candidate

Quality Metrics

$$Quality(t,c,R) = Similarity(t,c) * RelDiversity(c,R)$$

$$RelDiversity\ (c,R) = 1 \text{ if } R = \{\};$$

$$= \frac{\sum_{i=1..m}\left(1 - Similarity\left(c, r_i\right)\right)}{m}, \ otherwise$$

$$Quality(t,c,R)=(1-\ \alpha) * Similarity(t,c)+\alpha * RelDiversity(c,R)$$

$$Quality(t,c,R) = 2 \Big/ \left(\frac{1}{Similarity\ (t,c)} + \frac{1}{RelDiversity\ \ (c,R)} \right)$$

Diversity 2 – Greedy Selection

```
define GreedySelection (t,C,k)
begin
    R={}
    For i = 1 to k
        Sort C by Quality(t,c,R) for each c in C
        R=R + First (C)
        C=C – First (C)
    EndFor
return R
end
```

In practice one will not sort…

Diversity 3 – Bounded Greedy Selection

```
define **BoundedGreedySelection** (t,C,k)
Begin
    $C'$ = bk cases in C that are most similar to t
    R={}
    For i = 1 to k
        Sort $C'$ by Quality(t,c,R) for each c in $C'$
        R=R + First ($C'$)
        $C'$ =$C'$– First ($C'$)
    EndFor
return R
end
```

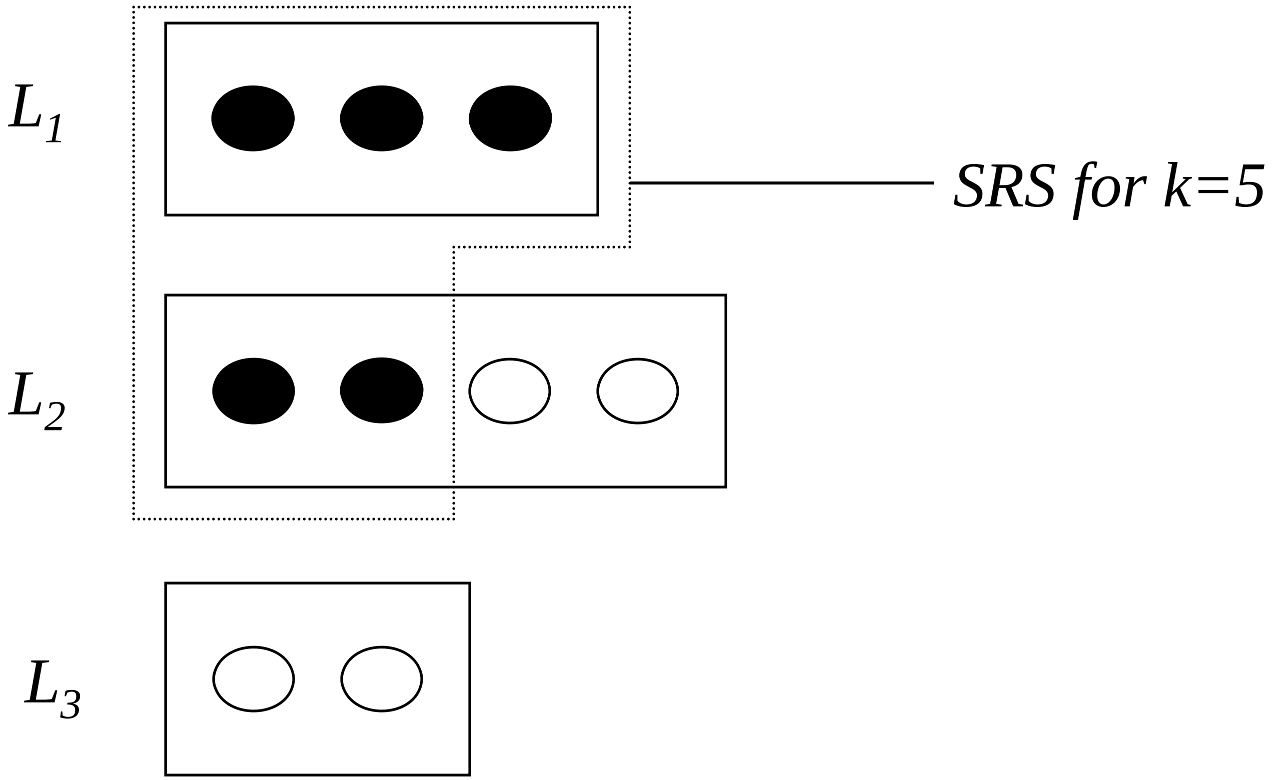
Similarity-Preserving Diversification (DCR-1)

- The process of constructing a retrieval set that is more diverse than SRS for a given query but no less similar is known as *similarity-preserving* diversification
- Maximally-similar retrieval set

  A retrieval set of the same size as of the SRS is maximally similar to the target query if it has the same average similarity to the target query as the SRS
- Similarity layers

  A target query *Q* partitions the case library into subsets $L_1, L_2, \ldots, L_n$ such that, for $1 \leq i \leq n-1$, all cases in L_i are equally similar to *Q*, and more similar to *Q* than any case in L_{i+1}. We refer to $L_1, L_2, \ldots, L_n$ as the similarity layers associated with *Q*.

Example of Similarity Layers



Theorem 1. *A maximally-similar retrieval set for a given query can differ from their SRS only in the cases it includes from the lowest similarity layer that contributes to the SRS.*

Maximizing Diversity

Without *any* concern for Similarity!

```
algorithm MaxD(Initial, RetrievalSet, Candidates, k)
begin
    $RetrievalSet \leftarrow Initial$
    while $|RetrievalSet| < k$ do
    begin
        $C_{best} \leftarrow first(Candidates)$
        $D_{max} \leftarrow relative\_diversity(C_{best}, RetrievalSet)$
        for all $C \in Candidates$ do
        if $relative\_diversity(C, RetrievalSet) > D_{max}$
          then begin
            $C_{best} \leftarrow \mathrm{C}$
            $D_{max} \leftarrow relative\_diversity(C, RetrievalSet)$
          end
        $RetrievalSet \leftarrow \{C_{best}\} \cup RetrievalSet$
        $Candidates \leftarrow Candidates - \{C_{best}\}$
    end
end
```

DCR-1 Algorithm

1. Given target query Q and required size k for retrieval set, it constructs SRS
2. Identifies the lowest similarity layer L_x
3. Let C_{max} be the case that is most similar to the target query
4. Calls MaxD with an initial retrieval set and a set of candidate cases that depend on whether $C_{max} \in L_x$
5. If $C_{max} \in L_x$, Initial$\leftarrow\{C_{max}\}$ and Candidates$\leftarrow L_x$-$\{C_{max}\}$
6. If $C_{max} \notin L_x$, Initial$\leftarrow$all cases in the layers above L_x (including C_{max}) and Candidates$\leftarrow L_x$

DCR-1 Algorithm

- DCR-1’s ability to increase diversity without loss of similarity depends on the underlying similarity measure
- With a more *fine-grained* similarity measure (than Sim_{MF}), opportunities for similarity-preserving diversification are *likely to occur less frequently*
- A reasonable strategy would be to round off the similarity values produced by a fine-grained similarity measure

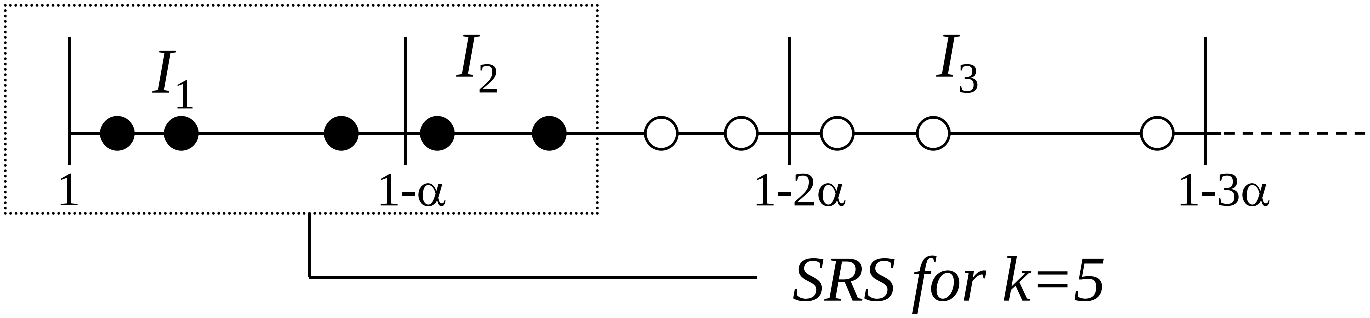
Similarity-Protected Diversification (DCR-2)

- Offers a compromise between the extremes of
  - insisting that similarity is fully preserved and
  - tolerating arbitrary loses in similarity
- Objective is to construct a retrieval set that is more diverse than the SRS
- Ensures that the loss of similarity is less than a predefined threshold value
- Uses a notion of Similarity *intervals*

Similarity Intervals

- Assume that the similarity measure *Sim* on which retrieval is based is such that $0<Sim(c,Q)<1$ for any case *c* and target query *Q*.
- Given a positive integer *r*, a target query *Q* partitions the set of cases with non-zero similarity into *similarity intervals* $I_1, I_2, .., I_r$ of width $\alpha = 1/r$. that is, for $1 \leq n \leq r$,

$$I_n = \{c: 1-n\alpha < Sim(c,Q) < 1-(n-1)\alpha\}$$



DCR-2 Algorithm

1. Given target query Q and required size k for retrieval set, it constructs SRS
2. Identifies the rightmost similarity interval layer I_x
3. Let C_{max} be the case that is most similar to the target query
4. Calls MaxD with an initial retrieval set and a set of candidate cases that depend on whether $C_{max} \in I_x$
5. If $C_{max} \in I_x$, Initial$\leftarrow\{C_{max}\}$ and Candidates$\leftarrow I_x$-$\{C_{max}\}$
6. If $C_{max} \notin I_x$, Initial$\leftarrow$all cases in the similarity intervals left to I_x (including C_{max}) and Candidates$\leftarrow I_x$

Theorems Related to DCR-2

Theorem2. *In* DCR-2, *the loss of average similarity relative to the SRS is always less than* α, *the width of the similarity intervals on which retrieval is based.*

Proof. Let *S* be the average similarity of the cases in the similarity intervals, if any, to the left of I_x.Let $s_1,s_2,..,s_m$ *and* $s'_1,s'_2,\ldots,s'_m$ be the similarities of the cases form I_x that contribute to SRS and DCR-2.

$$similarity(SRS) - similaity(DCR-2) = \frac{(k-m)S + \sum_{i=1}^{m} s_i}{k} - \frac{(k-m)S + \sum_{i=1}^{m} s'_i}{k}$$

$$= \frac{\sum_{i=1}^{m}\left(s_i - s'_i\right)}{k} < \frac{m\alpha}{k} \le \alpha$$

Theorem3. *In a recommender with* $\mathrm{Sim}_{\mathrm{MF}}$ *as the similarity measure,* DCR-1 *is equivalent to* DCR-2 *with* α*=1/r, where r is the number of case attributes on which retrieval is based.*

DCR-2 Algorithm

- DCR-2 limits the impact of diversification on average similarity using the above strategy
- An easier way to achieve this would be to insist on a minimum level of similarity to the target query among the retrieved cases
- One limitation of this approach is that depending on the required level of similarity, there may not be enough eligible cases to fill the retrieval set

Experimental Results (DCR-1)

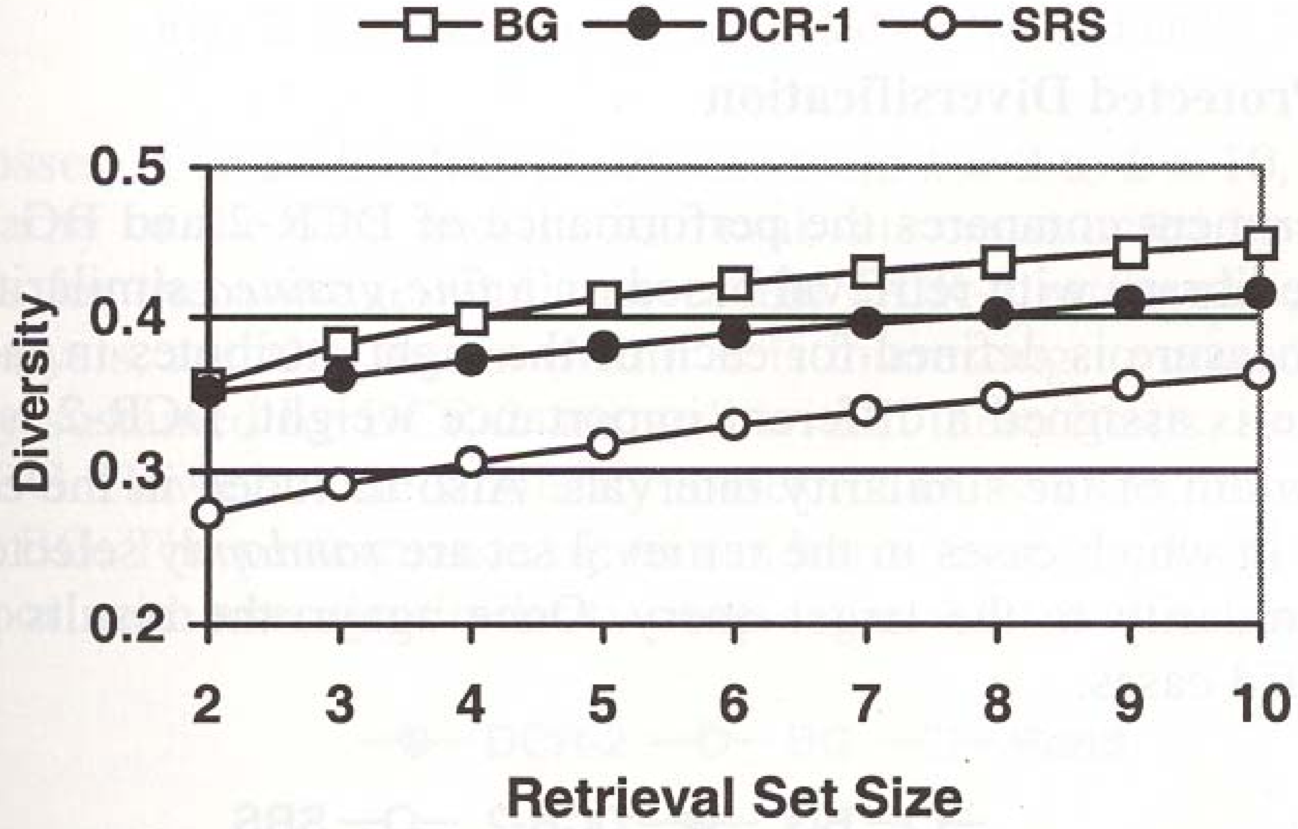


Fig. 4. Diversity gains provided by BG and DCR-1

Experimental Results (DCR-1)

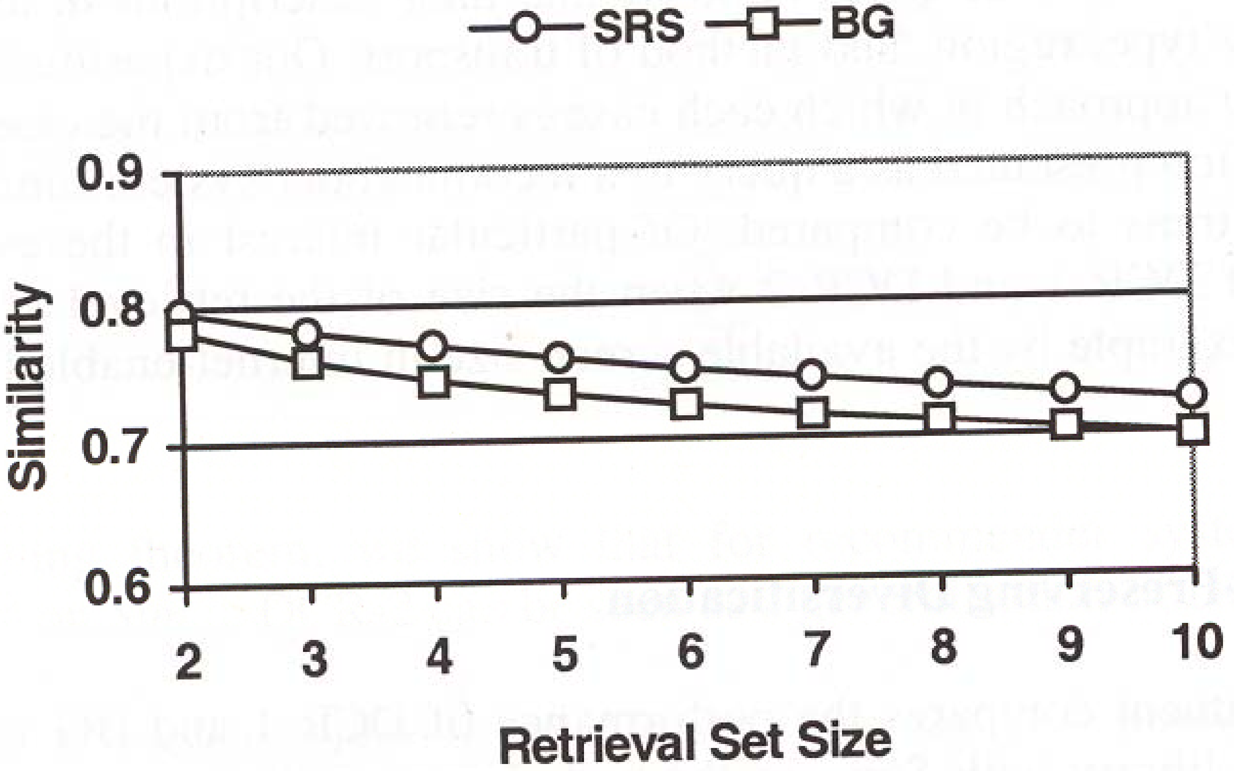


Fig. 5. Similarity losses sustained by BG

Experimental Results (DCR-2)

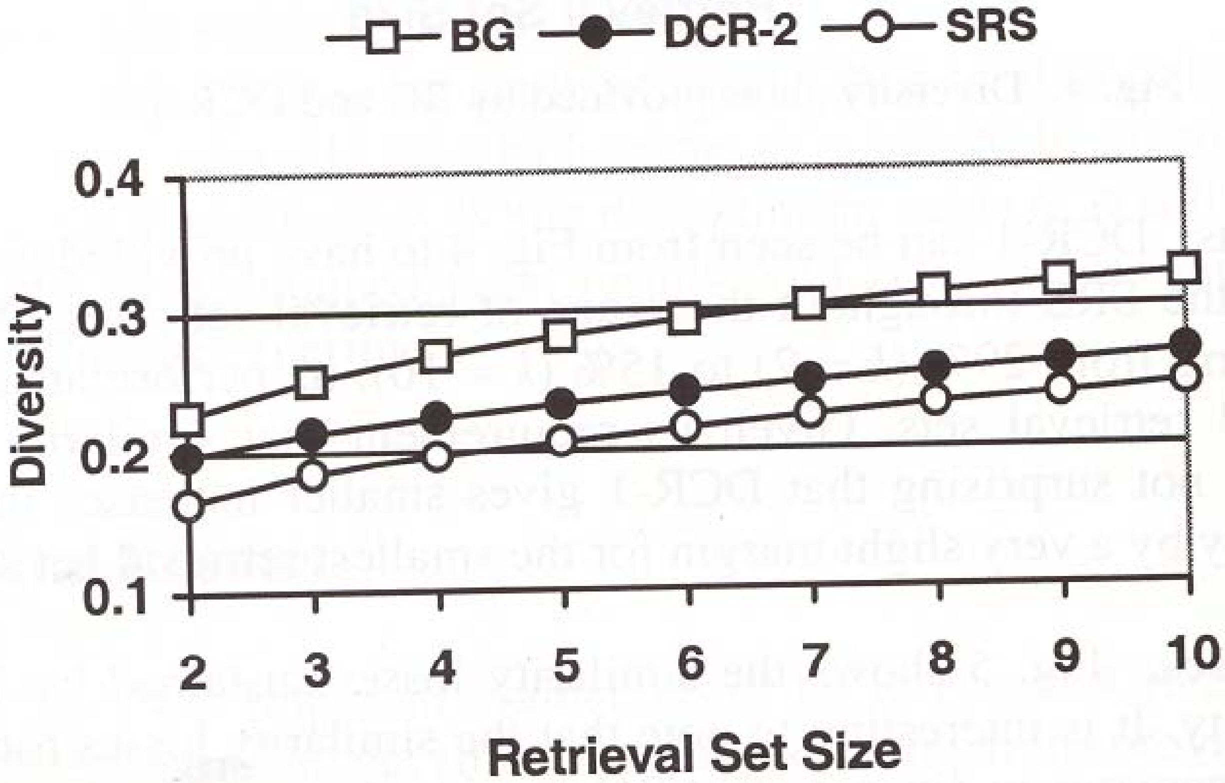


Fig. 6. Diversity gains provided by BG and DCR-2

Experimental Results (DCR-2)

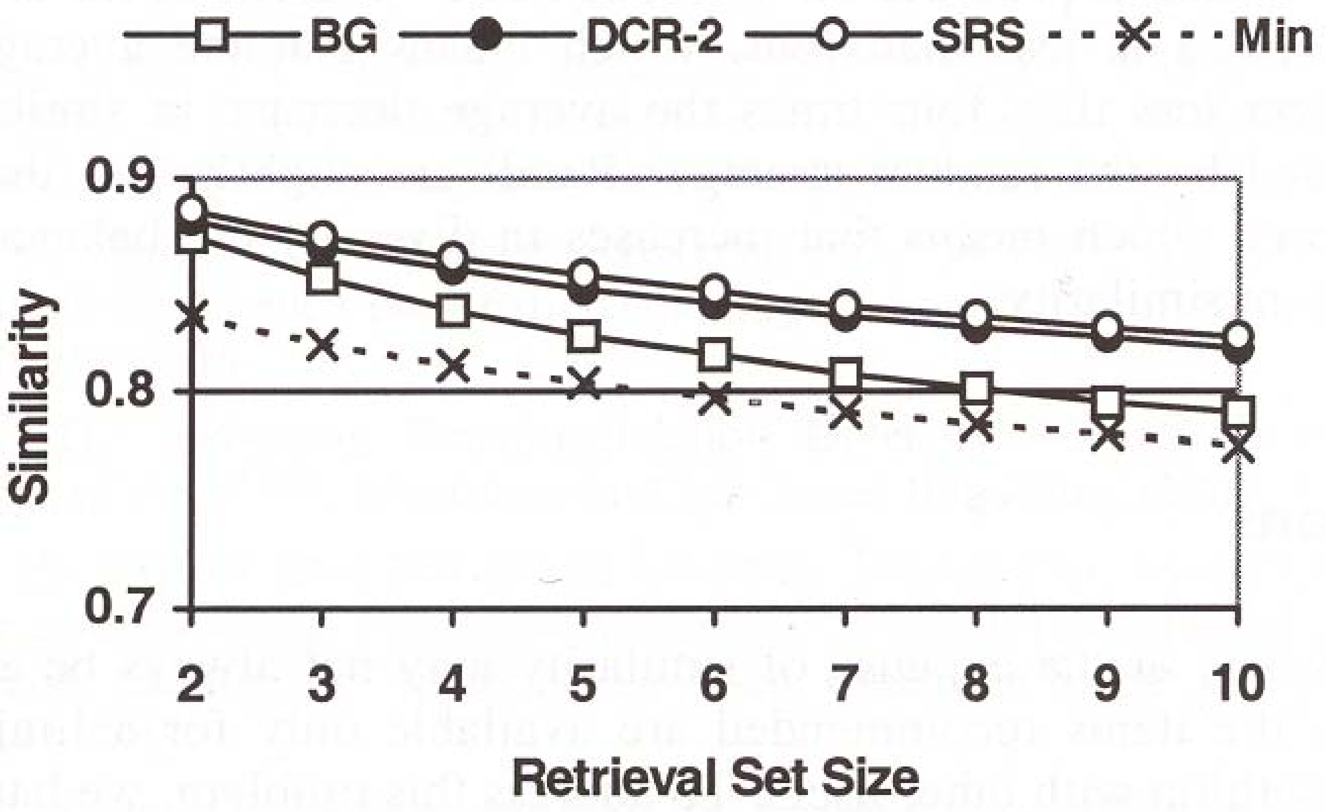


Fig. 7. Similarity losses sustained by BG and DCR-2

Experimental Results (DCR-2)

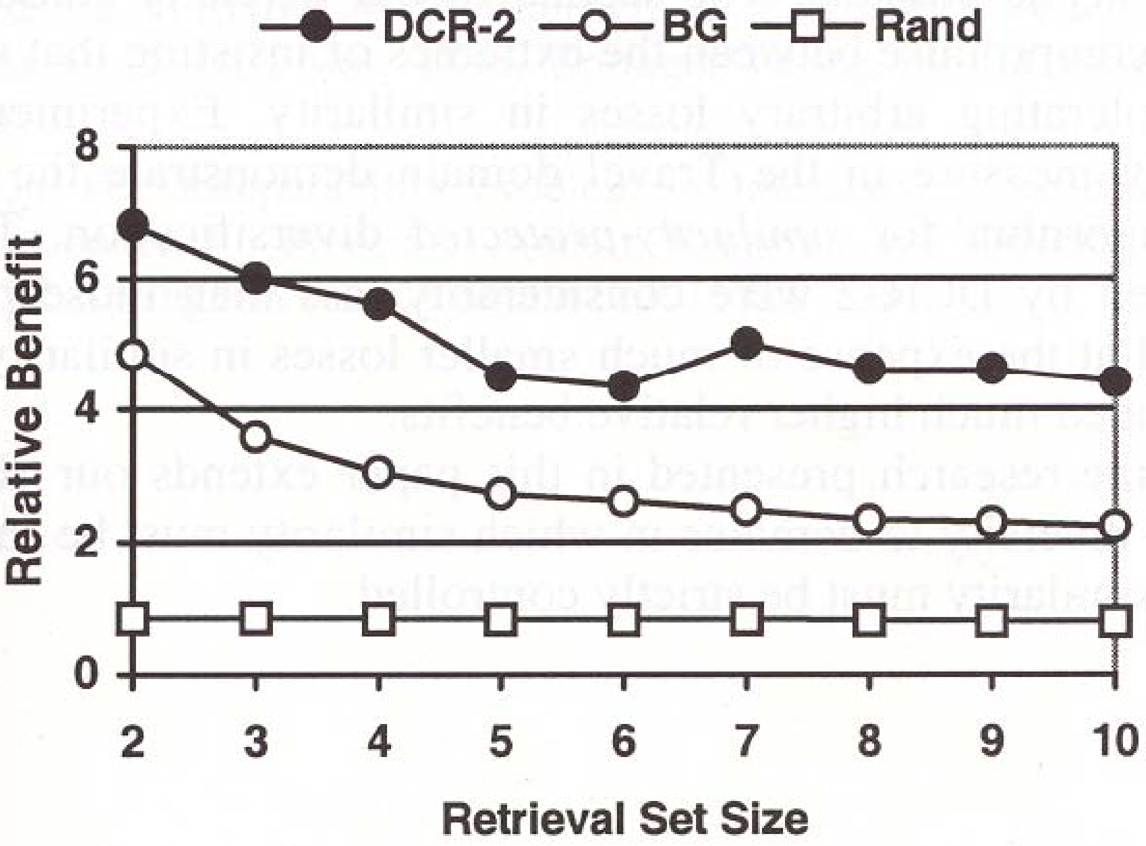


Fig. 8. Relative benefits provided by BG, DCR-2, and Rand

Relative benefit is the increase in diversity relative to SRS divided by the decrease in similarity

Conclusions

- Increasing diversity at the expense of similarity may not always be acceptable
- DCR-1 attempts to increase recommendation diversity while ensuring that similarity is fully preserved.
- DCR-2 ensures that the loss of similarity is less than a predefined threshold value

References

- David Mcsherry.: Diversity Conscious Retrieval. Proceedings of the Sixth European Conference, ECCBR 2002,Scotland 219-233

- Bradley, K., Smyth, B.: Improving Recommendation Diversity. Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland (2001) 85-94