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 most similar to the target query.
Measures of Similarity and Diversity

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

\[
Sim_{MF}(C, Q) = \frac{\left| \left\{ a \in A_Q : \prod a(C) = \prod a(Q) \right\} \right|}{|A_Q|}
\]

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

Note: Boolean Unweighted Match

Matching Features
Increasing Diversity

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

- Select next case $C$ such that its relative diversity w.r.t. $R = \{C_1, C_2, \ldots, C_n\}$ is the highest.

$$\text{relative
diversity}(C, R) = \sum_{i=1}^{n} \left(1 - \text{Sim}(C, C_i)\right)$$
## Example Case Library

<table>
<thead>
<tr>
<th>Case No.</th>
<th>beds = 4</th>
<th>Style = det</th>
<th>Rec</th>
<th>Loc = A</th>
<th>Sim$_{MF}$</th>
<th>SRS</th>
<th>BG</th>
<th>DCR-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>4</td>
<td>det</td>
<td>2</td>
<td>A</td>
<td>1.00</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>det</td>
<td>2</td>
<td>A</td>
<td>1.00</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>2</td>
<td>det</td>
<td>2</td>
<td>A</td>
<td>0.67</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>3</td>
<td>det</td>
<td>2</td>
<td>A</td>
<td>0.67</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>38</td>
<td>3</td>
<td>det</td>
<td>1</td>
<td>A</td>
<td>0.67</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>4</td>
<td>sem</td>
<td>2</td>
<td>A</td>
<td>0.67</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>det</td>
<td>1</td>
<td>C</td>
<td>0.67</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>sem</td>
<td>2</td>
<td>A</td>
<td>0.67</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>sem</td>
<td>2</td>
<td>D</td>
<td>0.33</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>49</td>
<td>2</td>
<td>det</td>
<td>2</td>
<td>C</td>
<td>0.33</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Similarity:** 0.80 0.67 0.80  

**Diversity:** 0.26 0.60 0.40
Example (contd..)

\[
similarity(BG) = \frac{1.00 + 0.67 + 0.67 + 0.67 + 0.33}{5} = 0.67
\]

\[
diversity(BG) = \frac{(0.33 + 0.33 + 0.33 + 0.67) + (0.33 + 0.67 + 1) + (0.67 + 1) + (0.67)}{5 \times \frac{(5-1)}{2}} = 0.60
\]

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

\(t: \text{target query}, \ C: \text{case-base}, \ k: \text{# results}, \ b: \text{bound}\)

define \textbf{BoundedRandomSelection} (t, C, k, b)

begin

\(C|a = bk \text{ cases in } C \text{ that are most similar to } t\)

\(R = k \text{ random cases from } C|a\)

return R

end

Retrieve \(b\) times more cases
Choose \(k\) from them randomly
Quality Metrics

\[ Quality(t,c,R) = Similarity(t,c) \times RelDiversity(c,R) \]

\[ RelDiversity(c,R) = 1 \text{ if } R = \{}; \]
\[ = \sum_{i=1..m} (1 - Similarity(c,r_i)) \text{, otherwise} \]

\[ Quality(t,c,R) = (1 - \alpha) \times Similarity(t,c) + \alpha \times RelDiversity(c,R) \]

\[ Quality(t,c,R) = 2 \sqrt{\left( \frac{1}{Similarity(t,c)} + \frac{1}{RelDiversity(c,R)} \right)} \]
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
define **BoundedGreedySelection** (t,C,k)

Begin

\( C' = \text{bk cases in C that are most similar to } t \)

R=\{

For \( i = 1 \) to \( k \)

Sort \( C' \) by Quality\((t,c,R)\) for each \( c \) in \( C' \)

R=R + First (\( C' \))

\( C' = C' – \text{First} (C') \)

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 $Q$. 
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

**algorithm** MaxD(Initial, RetrievalSet, Candidates, k)

**begin**

RetrievalSet ← Initial

**while** ⌊RetrievalSet⌋ < k **do**

**begin**

C\textsubscript{best} ← first(Candidates)

D\textsubscript{max} ← relative\_diversity(C\textsubscript{best}, RetrievalSet)

**for all** C ∈ Candidates **do**

**if** relative\_diversity(C, RetrievalSet) > D\textsubscript{max} **then begin**

C\textsubscript{best} ← C

D\textsubscript{max} ← relative\_diversity(C, RetrievalSet)

**end**

RetrievalSet ← \{C\textsubscript{best}\} ∪ RetrievalSet

Candidates ← Candidates - \{C\textsubscript{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* $I_1, I_2, \ldots, I_r$ of width $\alpha = 1/r$. That is, for $1 \leq n \leq r$,

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

![Diagram showing similarity intervals](attachment:image.png)

*SRS for $k=5$*
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

**Theorem 2.** In DCR-2, the loss of average similarity relative to the SRS is always less than $\alpha$, 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 $s_1, s_2, \ldots, s_m$ and $s'_1, s'_2, \ldots, s'_m$ be the similarities of the cases form $I_x$ that contribute to SRS and DCR-2.

\[
\text{similarity}(\text{SRS}) - \text{similarity}(\text{DCR-2}) = \frac{(k-m)S + \sum_{i=1}^{m} s_i}{k} - \frac{(k-m)S + \sum_{i=1}^{m} s'_i}{k} = \frac{\sum_{i=1}^{m} (s_i - s'_i)}{k} \leq \frac{m\alpha}{k} \leq \alpha
\]

**Theorem 3.** In a recommender with $\text{Sim}_{\text{MF}}$ as the similarity measure, DCR-1 is equivalent to DCR-2 with $\alpha = 1/r$, 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)

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

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
