Coverage-Optimized Retrieval

David McSherry
School of Computing and Information Engineering
University of Ulster, Coleraine BT52 1SA, Northern Ireland
dmg.mcsherry@ulster.ac.uk

Abstract

We present a generalization of similarity-based retrieval in recommender systems which ensures that for any case that is acceptable to the user, the retrieval set contains a case that is at least as good in an objective sense and so also likely to be acceptable. Our approach recognizes that similarity to the target query is only one of several possible criteria according to which a given case might be considered at least as good as another.

1 Introduction

An advantage of case-based reasoning (CBR) in product recommendation is that if none of the available products exactly matches the user's query, she can be shown the products that are most *similar* to her query [Wilke *et al.*, 1998]. However, a known limitation of similarity-based retrieval is that the most similar case may not be the one that is most acceptable to the user [Burkhard, 1998; Smyth and McClave, 2001]. The *k*-NN strategy of retrieving the *k* most similar cases only partially compensates for this limitation, as the number of cases that can be presented to the user is necessarily restricted in practice. So the existence of an acceptable case does not guarantee that it will be retrieved.

We present a new approach to retrieval called *coverage-optimized* retrieval (CORE) which ensures that for any case that is acceptable to the user, the retrieval set contains a case that is at least as good in an objective sense and so also likely to be acceptable. Similarity to the target query is only one of several possible criteria according to which a given case might be considered at least as good as another in the approach.

2 The CORE Retrieval Set

The similarity of a given case C to a target query Q over a subset A_Q of the case attributes A is typically defined as:

$$Sim(C, Q) = \sum_{a \in A_Q} w_a sim_a(C,Q)$$

where for each $a \in A_Q$, w_a is a numeric weight representing the importance of a and $sim_a(C, Q)$ is a measure of the similarity of $\pi_a(C)$, the value of a in C, to $\pi_a(Q)$, the preferred value of a. Apart from its similarity to the target query, another factor likely to influence the acceptability of a given case is the *compromises* it involves, or preferences of the user that it fails to satisfy [Burkhard, 1998]. Often in e-commerce domains, one can identify attributes whose values most users would prefer to maximize or minimize [Wilke *et al.*, 1998]. In CORE, we assume that the value specified by the user is a preferred minimum in the case of a *more-is-better* attribute or a preferred maximum in the case of a *less-is-better* attribute.

Definition 1 *For any query Q and case C we define:*

$$d_1(C, Q) = \{a \in A_Q : \ nominal(a), \pi_a(C) \neq \pi_a(Q)\}$$
$$d_2(C, Q) = \{a \in A_Q : \ more\text{-}is\text{-}better(a), \pi_a(C) < \pi_a(Q)\}$$
$$d_3(C, Q) = \{a \in A_Q : \ less\text{-}is\text{-}better(a), \pi_a(C) > \pi_a(Q)\}$$

$$d(C, Q) = \bigcup_{i = 1,2,3} d_i(C,Q)$$

So $d(C, Q)$ is the set of attributes with respect to which C fails to satisfy the user's preferences. Below we define four dominance criteria according to which a given case C_1 might be considered at least as good as another case C_2:

D0: $Sim(C_1, Q) \geq Sim(C_2, Q)$
D1: $Sim(C_1, Q) \geq Sim(C_2, Q)$ and $|d(C_1, Q)| \leq |d(C_2, Q)|$
D2: $Sim(C_1, Q) \geq Sim(C_2, Q)$ and $d(C_1, Q) \subseteq d(C_2, Q)$
D3: $sim_a(C_1, Q) \geq sim_a(C_2, Q)$ for all $a \in A_Q$

For example, if $sim_a(C_1, Q) \geq sim_a(C_2, Q)$ for all $a \in A_Q$ we say that C_1 dominates C_2 with respect to D3. We say that a given case C_2 is *covered* by a retrieval set RS if $C_2 \in RS$ or there exists $C_1 \in RS$ such that C_1 dominates C_2. The importance of coverage in this sense is that if an acceptable case that is not retrieved is covered by a retrieved case, then the retrieved case is also likely be acceptable. Another basic premise in our approach is that the likelihood of the retrieved case also being acceptable increases with the strength of the dominance criterion with respect to which it

dominates the acceptable case. For example, dominance of the acceptable case with respect to D0 (the usual similarity criterion) may not be enough to ensure that the retrieved case is also acceptable. However, if it dominates the acceptable case with respect to D3, then there is no attribute with respect to which it is less similar to the user's query.

In *k*-NN, all cases are dominated by the most similar case with respect to D0; so even 1-NN provides full coverage of the case library with respect to D0, the weakest of our dominance criteria. The aim in CORE is to construct a retrieval set that provides full coverage of the case library with respect to any of the dominance criteria we have identified. In Figure 1, Q is the target query and *Candidates* is a list of candidate cases for addition to the retrieval set *RS*. We assume that the candidate cases, initially all cases in the case library, are sorted in order of non-increasing similarity, and that if C_1, C_2 are equally similar cases such that C_1 dominates C_2 but C_2 does not dominate C_1, then C_1 is listed before C_2 in *Candidates*.

```
algorithm CORE(Q, Candidates, RS)
begin
  RS ← ϕ
  while |Candidates| > 0 do
  begin
     C1 ← first(Candidates)
     RS ← {C1} ∪ RS
     cover(C1) ← {C1}
     for all C2 ∈ rest(Candidates) do
     if C1 dominates C2
     then cover(C1) ← cover(C1) ∪ {C2}
     Candidates ← Candidates - cover(C1)
  end
end
```

Figure 1. Generic algorithm for coverage-optimized retrieval.

We refer to the versions of CORE based on D1, D2 and D3 as CORE-1, CORE-2 and CORE-3 respectively. It is worth noting that with D0 as the dominance criterion, CORE is equivalent to 1-NN, and so CORE is in fact a generalization of similarity-based retrieval.

Theorem 1 *The* CORE *retrieval set provides full coverage of the case library and no smaller retrieval set can provide full coverage of the case library.*

While it can easily be shown that the maximum possible sizes of the CORE-1 and CORE-2 retrieval sets for a given query Q are $|A_Q| + 1$ and $2^{|A_Q|}$ respectively, the maximum possible size of the CORE-3 retrieval set is not as easily determined. In practice, CORE-1 and CORE-2 retrieval sets are usually much smaller than their maximum possible sizes. Figure 2 shows the maximum, average, and minimum sizes of CORE retrieval sets for full-length queries on the Travel case library (*www.ai-cbr.org*). As might be expected, there is a trade-off between the strength of the dominance criterion in terms of which coverage is defined and the size of the retrieval set required to provide full coverage of the case library. Though unable to compete with CORE-1 in terms of coverage efficiency, CORE-2 has an average retrieval-set size of only 7.5 cases.

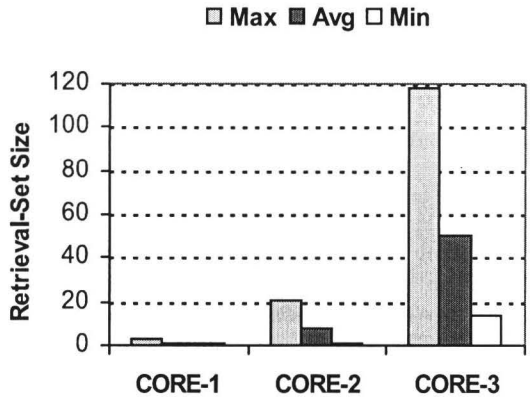

Figure 2. Size of the CORE retrieval set for full-length queries on the Travel case library.

3 Conclusions

Coverage-optimized retrieval (CORE) is a generalization of similarity-based retrieval which ensures that for any case that is acceptable to the user, the retrieval set contains a case that is at least as good according to a given dominance criterion. Our empirical results show that the size of the retrieval set required to provided full coverage of the case library in this sense increases with the strength of the dominance criterion in terms of which coverage is defined. However, CORE-2 offers a good compromise between strength of dominance and retrieval-set size, with an average of only 7.5 cases required to provide full coverage for full-length queries on a case library containing over 1,000 cases.

References

[Burkhard, 1998] H.-D. Burkhard. Extending some concepts of CBR - foundations of case retrieval nets. In M. Lenz, B. Bartsch-Spörl, H.-D. Burkhard and S. Wess (eds.), *Case-Based Reasoning Technology*, pages 17-50. Berlin-Heidelberg, Springer-Verlag, 1998.

[Smyth and McClave, 2001] B. Smyth and P. McClave. Similarity vs. diversity. In *Proceedings of the Fourth International Conference on Case-Based Reasoning*, pages 347-361, Vancouver, Canada, 2001. Springer-Verlag.

[Wilke *et al.*, 1998] W. Wilke, M. Lenz and S. Wess. Intelligent sales support with CBR. In M. Lenz, B. Bartsch-Spörl, H.-D. Burkhard and S. Wess (eds.), *Case-Based Reasoning Technology*, pages 91-113. Berlin-Heidelberg, Springer-Verlag, 1998.