Feature Selection for Ranking

Xiubo Geng, Tie-Yan Liu, Tao Qin, Hang Li.
SIGIR'07

Presenter: Suhan Yu

Introduction

- Traditionally only a small number of strong features were used to represent relevance and to rank documents.
- In recent years, with the development of the supervised learning algorithms like Ranking SVM and RankNet, it becomes possible to incorporate more features (strong or weak) into ranking models.
- Feature selection can help enhance accuracy in many machine learning problems.
- Feature selection can also help improve the efficiency of training.

Feature selection method

- Suppose the goal is to select $t \ (1 \leq t \leq m)$ features from the entire feature set $\{v_1, v_2, \ldots, v_m\}$.
- Assign an importance score to each feature.
  - MAP
  - NDCG
  - Loss function
- Similarity between features
  - Kendall's τ

$$\tau_q(v_i, v_j) = \frac{\#\{(d_s, d_t) \in D_q \mid d_s \prec_{v_i} d_t \ \ and \ \ d_s \prec_{v_j} d_t\}}{\#\{(d_s, d_t) \in D_q\}}$$

$d_s \prec_{v_i} d_t$ implies that instance d_t

is ranked ahead of instance d_s by feature v_i

Optimization formulation

$$\max \sum_i w_i x_i$$

$$\min \sum_i \sum_{j \neq i} e_{i,j} x_i x_j \qquad e_{i,j} = \tau_{i,j}$$

$$x_i \in \{0,1\} \quad i = 1, \ldots, m$$

$$\sum_i x_i = t \qquad t \text{ denotes the number of select features}$$

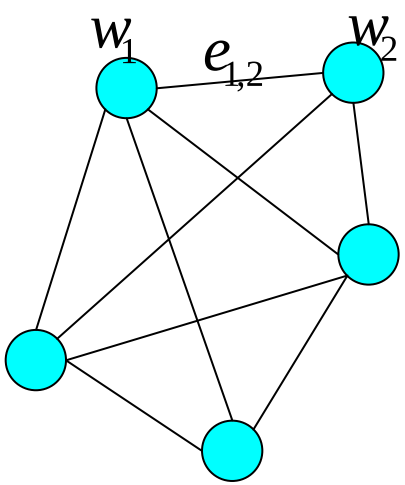
- Maximize the total important scores and minimize the total similarity scores.
- We take a common approach in optimization and convert multi-objective programming to single-objective programming using linear combination.

$$\max \sum_i w_i x_i - c \sum_i \sum_{j \neq i} e_{i,j} x_i x_j$$

c is a parameter to balance the two objectives .

Solution to optimization problem

- Greedy search



Algorithm GAS (Greedy search Algorithm of feature Selection)

1. Construct an undirected graph G_0, in which each node represents a feature, the weight of node v_i is w_i and the weight of an edge between node v_i and node v_j is $e_{i,j}$.

2. Construct a set S to contain the selected features. Initially $S_0 = \varnothing$.

3. For $i = 1 \ldots t$,

   (1) Select the node with the largest weight, without loss of generality, suppose that the selected node is v_{k_i}.

   (2) A punishment is conducted on all the other nodes according to their similarities with v_{k_i}. That is, the weights of all the other nodes are updated as follows.

   $$w_j \leftarrow w_j - e_{k_i,j} * 2c, \quad j \neq k_i$$

   (3) Add v_{k_i} to the set S and remove it from graph G together with all the edges connected to it:

   $$S_{i+1} = S_i \cup \{v_{k_i}\}, \quad G_{i+1} = G_i \setminus \{v_{k_i}\}$$

4. Output S_t.

Fig. 1 Greedy algorithm of feature selection for ranking

Solution to optimization problem

Proof:

The condition $S_{t+1} \supset S_t$ indicates that when selecting the (t+1)-th feature, we do not change the already-selected t features. Denote $S_t = \{v_{k_i} \mid i = 1, \ldots, t\}$, where v_{k_i} is the k_i-th feature selected in the i-th iteration. Then the task turns out to be that of finding the (t+1)-th feature so that the following objective can be met.

$$\max \sum_{i=1}^{t+1} w_{k_i} - c \sum_{i=1}^{t+1} \sum_{j \neq i} e_{k_i, k_j} \qquad (3)$$

Since $e_{k_i,k_j} = e_{k_j,k_i}$, we can rewrite (3) as

$$\max \sum_{i=1}^{t+1} w_{k_i} - 2c \sum_{i=1}^{t} \sum_{j=i+1}^{t+1} e_{k_i, k_j} \qquad (4)$$

And since $S_{t+1} \supset S_t$ and $S_t = \{v_{k_i} \mid i = 1, \ldots, t\}$, (4) equals

$$\max_s \{(\sum_{i=1}^{t} w_{k_i} - 2c \sum_{i=1}^{t-1} \sum_{j=i+1}^{t} e_{k_i, k_j}) + (w_s - 2c \sum_{i=1}^{t} e_{k_i, s})\}$$

Note that the first part of the objective is a constant with respect to s, and thus the goal becomes to select the node maximizing the second part. It is easy to see that in our greedy search algorithm, for the (t+1)-th iteration, the current weight for each node v_s is $(w_s - 2c \sum_{i=1}^{t} e_{k_i, s})$. Therefore, selecting the node with the largest weight is equivalent to selecting the feature that satisfies the optimization requirements in (2).■

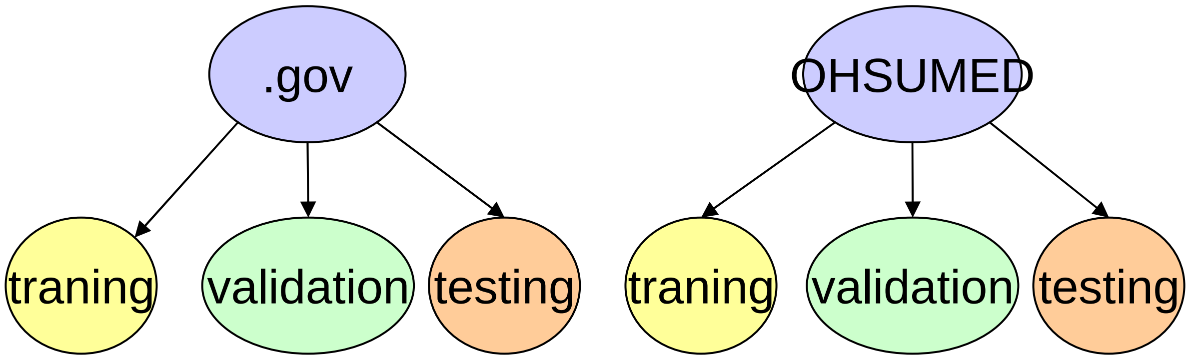
$$\begin{matrix} e_{k1,k1} & e_{k1,k2} & e_{k1,k3} & e_{k1,k4} \\ e_{k2,k1} & e_{k2,k2} & e_{k2,k3} & e_{k2,k4} \\ e_{k3,k1} & e_{k3,k2} & e_{k3,k3} & e_{k3,k4} \\ e_{k4,k1} & e_{k4,k2} & e_{k4,k3} & e_{k4,k4} \end{matrix}$$

Experiment

- Datasets
  - .gov data
    - used in the topic distillation task of Web track of TREC 2004
    - There are in total 1,053,110 documents and 75 queries with binary relevance judgments in the dataset.
    - used the BM25 model to retrieve the top 1000 documents for each query.
    - extracted 44 features for each document
      - features like document length, term frequency, inverse document frequency, BM25, language model features, PageRank, and HITS, and newly-proposed features, such as HostRank and relevance propagation.

Experiment

- Datasets
  - OHSUMED data
    - used in many experiments in information retrieval, including the TREC-9 filtering track.
    - Bibliographical document collection.
    - There are in total 16,140 query-document pairs upon which three levels of relevance judgments are made: “definitely relevant”, “possibly relevant”, and “not relevant”.
    - extracted in total 26 features from each document.



Evaluation measure

- MAP
  - Mean average precision
  - It is assumed that there are two types of documents: positive and negative (relevant and irrelevant).

$$P(n) = \frac{number\ of\ positive\ ins\tan ce\ within\ top\ n}{n}$$

$$AP = \sum_{n=1}^{N} \frac{P(n) \times pos(n)}{\text{number of positive instance}}$$

  - the OHSUMED dataset has three types of labels. We define “definitely relevant” as *positive* and the other two as *negative* when calculating MAP.

Evaluation measure

- NDCG
  - Normalized discount cumulative gain

$$N(n) = Z_n \sum_{j=1}^{N} \frac{2^{R(j)} - 1}{\log(1 + j)}$$

$n : position$

$R(j)$ denotes score for rank j

Z_n is a normalization factor

- Proposed algorithm

Algorithm	Description
GAS-E	In GAS-E we use evaluation measures (e.g. NDCG, MAP) to calculate the importance score of each feature.
GAS-L	In GAS-L we use the empirical loss of ranking model to measure the importance of each feature. For example, in Ranking SVM, we use pair-wise ranking error; and in RankNet, we use the cross entropy loss.

- Information Gain (IG)
  - Measures the reduction in uncertainty (entropy) in classification prediction
- Chi-square (CHI)
  - Measures the degree of independence between the feature and the categories.

Chi-square

- Under the null hypothesis: (*jaguar* and *auto*-independent): How many co-occurrences of *jaguar* and *auto* do we expect?
  - We would have: $Pr(j,a) = Pr(j) \times Pr(a)$
  - So, there would be: $N \times Pr(j,a)$, i.e. $N \times Pr(j) \times Pr(a)$
    - $Pr(j) = (2+3)/N$
    - $Pr(a) = (2+500)/N$
    - *Where* $N = 2+3+500+9500$
  - Which is: $N \times (5/N) \times (502/N) = 2510/N = 2510/10005 \approx 0.25$

	Term = jaguar	***Term ≠ jaguar***
Class = auto	**2** ***(0.25)***	**500** ***(502)***
Class ≠ auto	**3** ***(4.75)***	**9500** ***(9498)***

expected: f_e

observed: f_o

$$\chi^2(j,a) = \sum \frac{(O-E)^2}{E} = \frac{(2-0.25)^2}{0.25} + \frac{(3-4.75)^2}{4.75} + \frac{(500-502)^2}{502} + \frac{(9500-9498)^2}{9498} = 12.9$$

Training Procedures

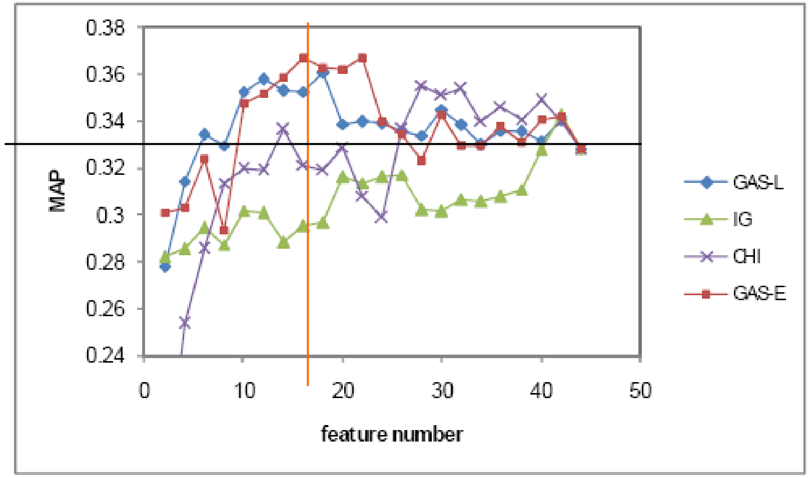
Feature selection (training set) (IG, CHI, Proposed algorithm)

Tuning Ranking model (validation set) (Ranking SVM, RankNet)

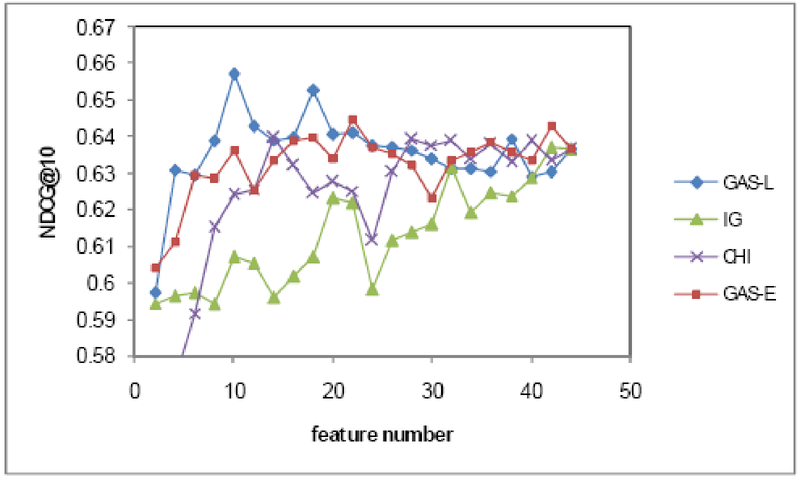
Tuning parameter c

Evaluation (test set) (MAP, NDCG)

Experimental Results

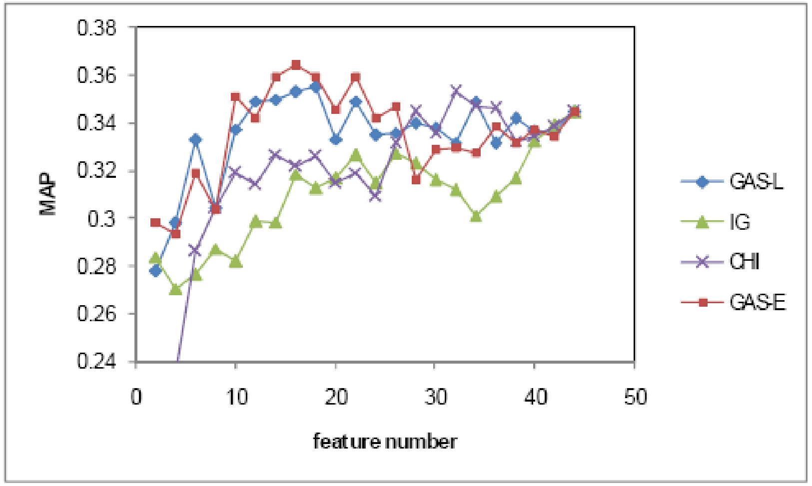


(a) MAP of Ranking SVM

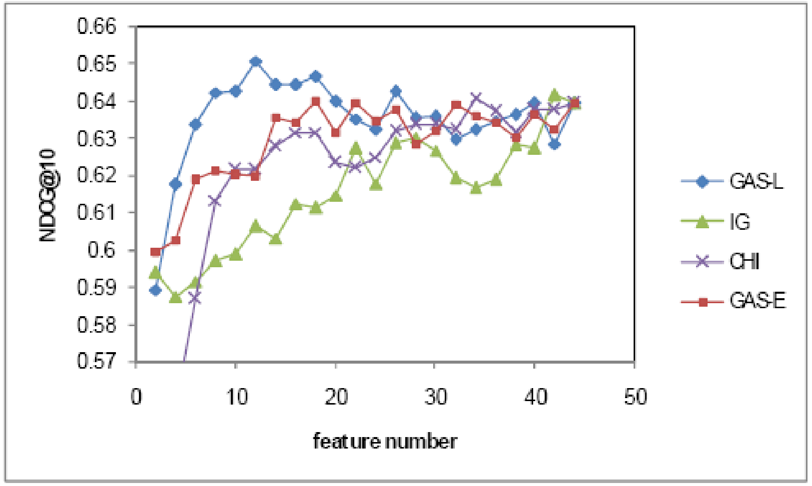


(b) NDCG@10 of Ranking SVM

Fig. 2 Ranking accuracy of Ranking SVM with different feature selection methods on the .gov dataset



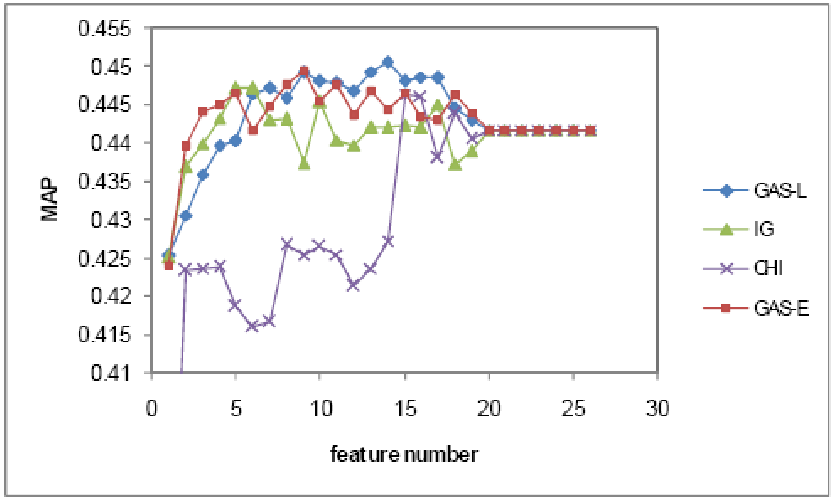
(a) MAP of RankNet



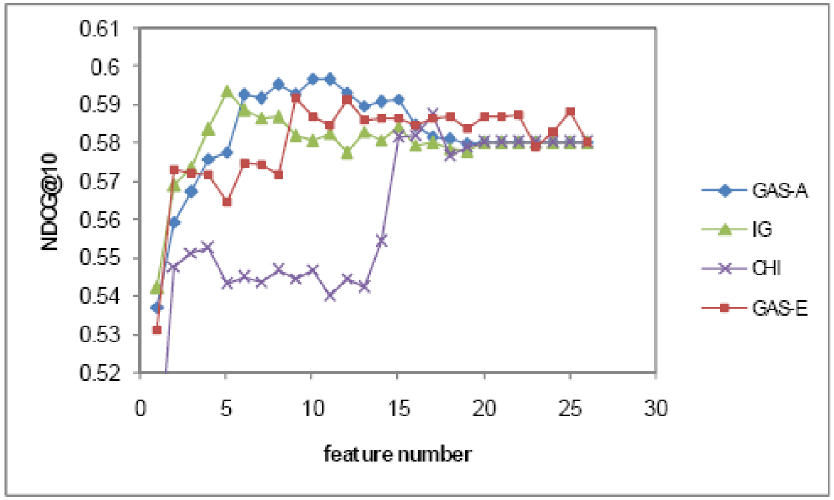
(b) NDCG@10 of RankNet

Fig. 3 Ranking accuracy of RankNet with different feature selection methods on the .gov dataset

Experimental Results

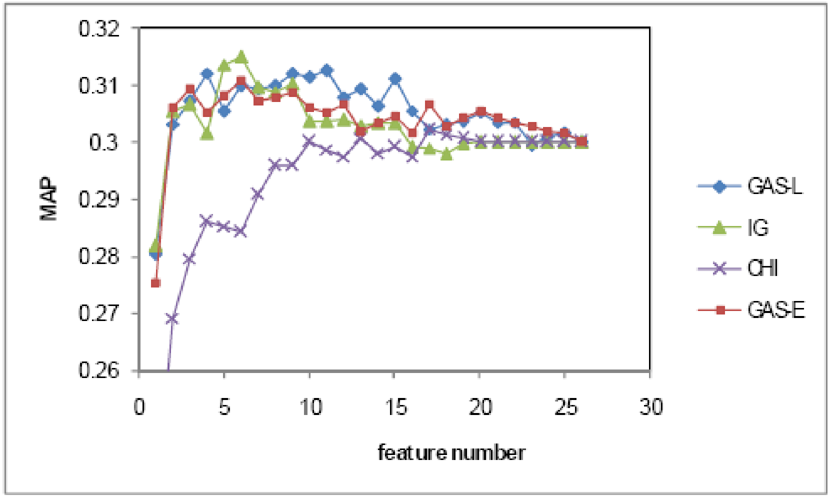


(a) MAP of Ranking SVM

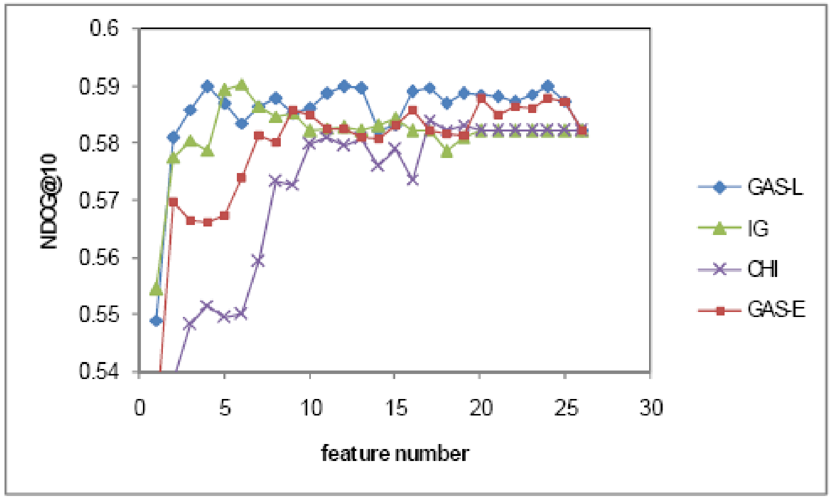


(b) NDCG@10 of Ranking SVM

Fig. 4 Ranking accuracy of Ranking SVM with different feature selection methods on the OHSUMED dataset



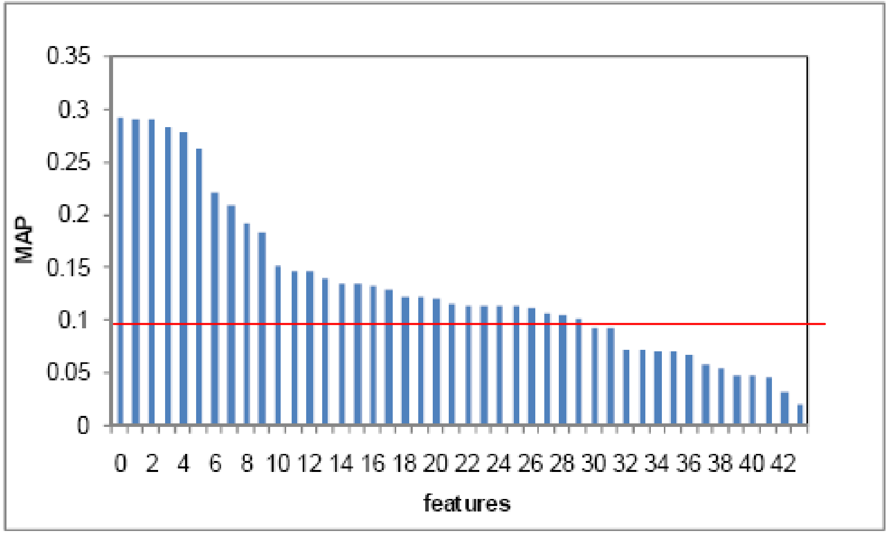
(a) MAP of RankNet



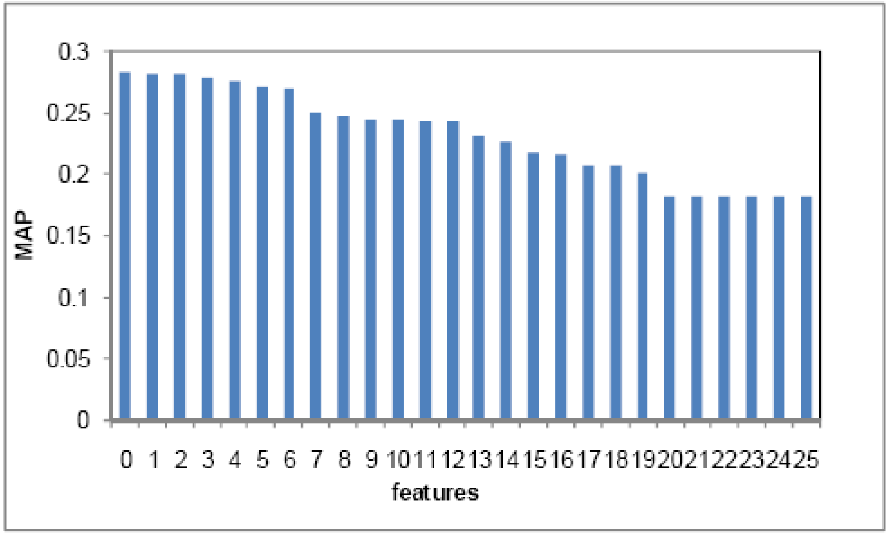
(b) NDCG@10 of RankNet

Fig. 5 Ranking accuracy of RankNet with different feature selection methods on the OHSUMED dataset

Experimental Results



(a) The .gov dataset



(b) The OHSUMED dataset

Fig. 6 MAP of individual features in the two datasets

Experimental Results

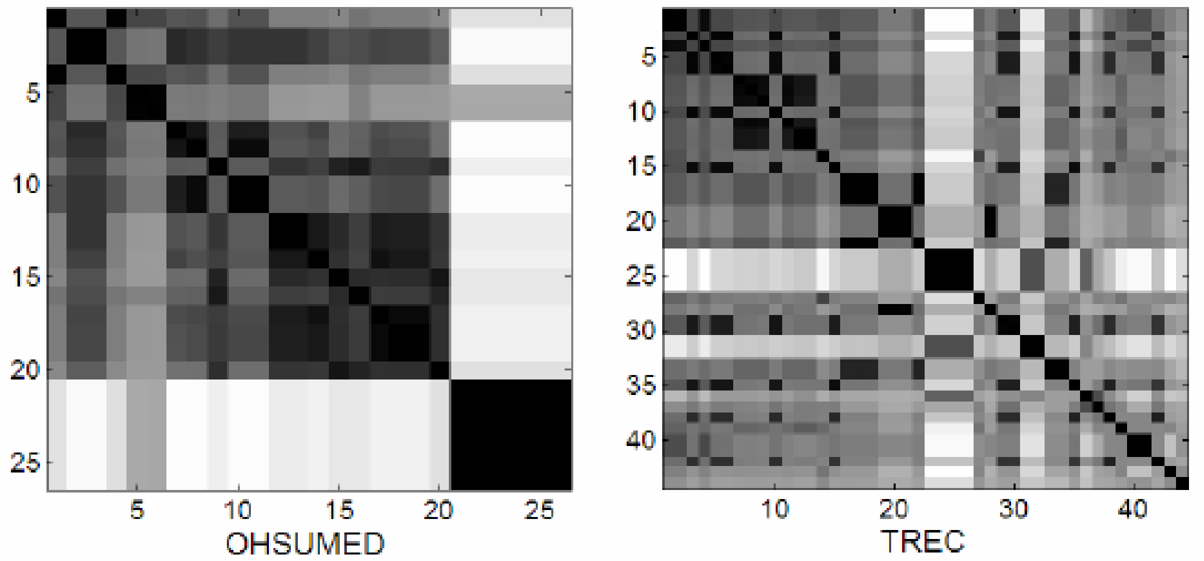


Fig. 7 Similarity between features in the two datasets

the OHSUMED dataset, there are only two
large blocks, with most features similar to each other. In this
case,the similarity punishment in our approach cannot work well.

Conclusion

- If the effects of features vary largely and there are redundant features, this method can work very well.
- There are two objectives in our optimization method for feature selection. In this paper combined them linearly for simplicity. In principle, one could employ other ways to represent the tradeoff between the two objectives.
- This paper have demonstrated the effectiveness with two datasets, and with a small number of manually extracted features. It is necessary to further conduct experiments on larger datasets and with more features.