**Query Expansion for Email Search**

Three state-of-the-art expansion methods are examined:

1) a global translation-based expansion model;

2) a personalized-based word embedding model;

3) the classical pseudo-relevance-feedback model.

Used two large datasets of real user queries and corresponding messages

the relevance of the message to the query is estimated by a complicated scoring function that considers many signals of relevance, including the message freshness, the textual similarity to the query, the user interaction with the message, and many more signals [2]

While the average query length on the Web is about three terms per query, the average length in the email domain is only 1.5 terms per query [2]

(problems with extremely short queries) -> Query expansion can help

**Translation model**

Given a query log of a commercial Web mail service, we extract a large dataset of email queries, each associated with clicked (presumably relevant) messages. e data pairs are used to train a translation model that maps query terms to relevant message terms. queries are then expanded by the most related terms to the query terms.

ignoring user personal preferences and biases

------------

Expands the query using a global translation model, learned from the system’s query log.

Query logs are composed from the user queries, each one is associated with a list of up to 100 messages, where messages that were clicked by the user are marked relevant. If there are more than one clicked, only the last one is relevant.

Query pairs < Q,s > where Q is original user query and “s” is the texto of the subject of the clicked message for that query.

IBM model 1 used for query expansion. For each term **t** the probability distribution over the Vocabulary V is Pr(w|t) = prior probability of translating term **t** to term **w**.

We score the vocabulary terms for a given query:

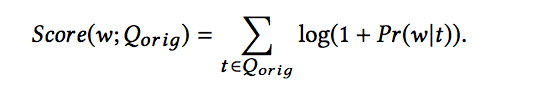


Ilustração 1

We then select the top-k scored words for query expansion. The selected terms are added to the original query, weighted according to their translation score, where term weight is normalized with respect to all k expanded terms.

Finnaly, the expanded query is linearly combined with the original query.

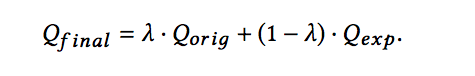


Ilustração 2

**Personalized model**

Our second expansion model expands the query in a personal manner.

The content of the personal mailbox of each user is used to train a word embedding model [11] which projects all terms of the messages into a dense lower dimensional space. e nearest neighbors of the query terms in the embedding domain are then used for expansion [8].

------------------------

As users search over their own data, using their own personal vocabulary, expanded terms should reflect their own preferences.

In the personal query expansion model, the query is expanded with terms that are “semantically similar” to the query terms, where similarity is measured in the context of the personal user content.

Specifically, we use the Word2Vec Continuous Bag- of-Words (CBOW) approach [11] which embeds terms in a vector space based on their co-occurrence in windows of text. e cosine similarity between the term vectors was shown to correlate with semantic similarity. Accordingly, we select terms similar to the query in this vector space for query expansion.

The “Subject” field terms, and the “From” and “To” field terms of all user mailbox messages are represented by the Word2Vec model.



Ilustração 3

The final term score for the given query is determined by aggregation over the query terms using Equation 1, and the top-k scored terms are then added to the query using Equation 2.

**Feedback model**

In contrast, relevance feedback based methods do not depend on any auxiliary data resource. Pseudo relevance feedback (PRF) methods construct an expansion model from the search results and then expand the query using the inferred model [9].

The expanded query can be used for re-ranking the search results, or to be submited as a new search task

PRF methods improve the search and efectiveness on average, however, they are very sensitive to the quality of the original search results.

-----------------

In this model, the top scored messages retrieved for the query are used for constructing an expansion model.

Given the list of retrieved messages, the relevance model is used to construct the expanded query.

model is defined by Pr(t|RM1) = m∈M PrMLE(t|m)Pr(m|q)

where PrMLE(t|m) is the maximum likelihood estimated of term t in message m

t f (t ∈ m) is the number of occur- rences of t in m.

A message is de ned here as a concatenation of its “Subject”, “From”, and “To” eld terms.

Given the relevance model, we select the top-k scored terms according to the model and expand the query with them using Equation 2. e expanded query is then used to re-rank the search results.

**Ranking model**

Explore the potential merit of the query expansion approaches over an effective representation of the original query. There were used two retrieval models.

First one measures only textual similarity of the query to the message based on the sequence dependence model. Each message is then scored with respect to the query using a linear interpolation of the SDM scores of its fields.

The second retrieval model was used a learning-to-rank scoring function for email search, which considers the message freshness, the textual similarity to the query, and the user actions on the messages. The similarity of the expanded query to the message is added as an additional feature to the scoring function.

**Experiments**

Each entry in the query log consists of a query text and the corresponding result list

we consider only search results that were received no more than six months before the query was issued, and only the latest clicked message in the result list is considered relevant.

we constructed a personal search index for each of the users in our training set

Messages were collected differently in two query sets.

* **LogData**, we collected the union of all messages of the user in the query log, i.e., all messages retrieved as a result of at least one of the user’s queries
* **MailBox**, we used the mail data of a small group of users for which their mailboxes are fully available for us; all of the messages in the specific user’s mail box constitute her personal collection.

Queries and messages were processed in the same manner. Stopwords were removed, and text was stemmed using Lucene’s minimal English stemmer1

For training the translation model we sampled queries of 11,200 users, over a period of 9 months, resulting in approximately 2.5M pairs of queries and subjects. these users are different from those whose data was collected for our query sets.

the performance of the different models is evaluated using two measures: Mean Reciprocal Rank(MRR) and success@n.

The models that we study incorporate several free parameters. Free parameter values were set in the following manner. We split the query set at random to training (1/3) and test (2/3) sets. Then, we choose parameter values that maximize MRR on the training set, and report performance on the test set.

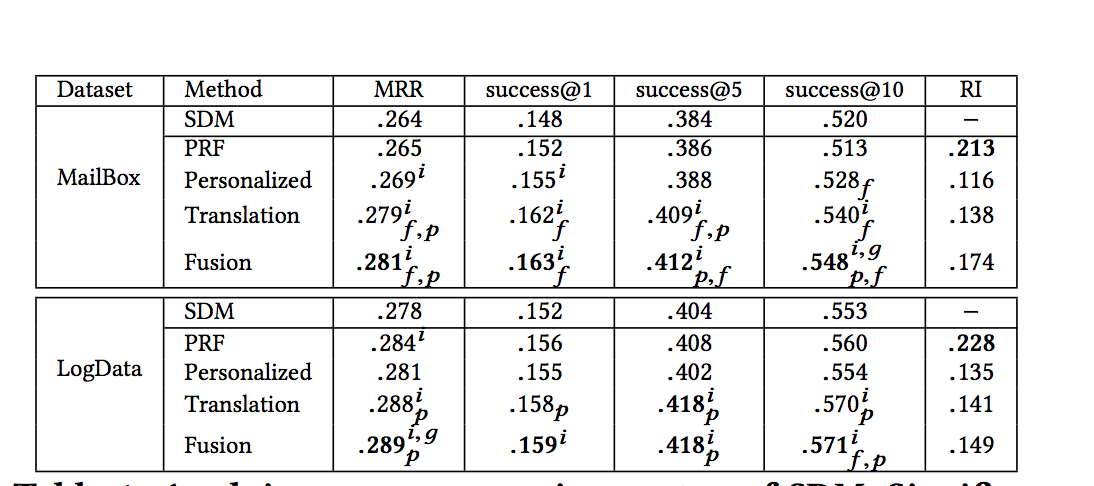
**Experimental Results**

According to the results, all query expansion approaches outperform SDM in a vast majority of reference comparisons; most of the MRR improvements are statistically significant.

We also study the Fusion approach which utilizes both the Personalized and the Translation models

Among the three suggested approaches, Translation is the most effective according to Table 1.

the performance of Translation (MRR and success@n) often dominates those of PRF and Personalized



The Personalized expansion approach is effective in the case of MailBox. It outperforms SDM for all evaluation measures .

In LogData, on the other hand, the Personalized approach never outperforms SDM in signicant manner. Tis difference in performance between the two datasets can be ascribed to the number of available messages per user which is larger in the case of MailBox. As the Word2Vec approach requires large amounts of data in order to learn an efective model, the resulted personalized models in the case of LogData may be less efective.

The PRF approach signi cantly improves SDM in terms of MRR only in LogData.

In addition, PRF is the most robust expansion approach in terms of RI. A possible explanation for this is that PRF models are induced from an initial result list. Such models, therefore, tend to be closer to the original query as compared to other approaches that utilize some external source (which is constructed in a non-query dependent manner). Consequently, PRF models may be less risky.

In Fusion, the term lists of both approaches are linearly interpolated using a free parameter learned from {0.0, 0.1, 0.5, 0.9, 1.0}. we fuse expansion term lists of 50 terms of Translation and Personalized, and then extract k terms with the highest combined score for expansion (using Equation 2). the performance of Fusion dominates those of the two approaches.

In terms of RI, we can see that Fusion is much more robust than both approaches. Given the high performance of the oracle, the question of how to optimally combine the two approaches remains open.