MUSETS: Diversity-aware Web Query Suggestions for Shortening User Sessions

M. Sydow[1,2], C. I. Muntean[3], F. M. Nardini[3],
S. Matwin[1,4], F. Silvestri[5]

Polish Academy of Sciences, Warsaw, Poland [1]
Polish-Japanese Institute of Information Technology, Warsaw, Poland [2]
ISTI-CNR, Pisa, Italy [3]
Big Data Institute, Dalhousie University, Halifax, Canada [4]
Yahoo Labs, London, UK [5]

ISMIS, Lyon, France
October 21-23, 2015

Generating search query suggestions triggered by an ambiguous or underspecified user query

- As an optimization problem
  - Given an ambiguous user query, the goal is to propose the user a set of query suggestions optimizing a set-wise objective function.
    - The function models the expected number of steps carried out by a user until reaching a satisfactory query formulation
    - The function is diversity-aware, as it naturally enforces high coverage of different alternative continuations of the user session
- For modeling the topics covered by the queries, we also use an extended query representation based on entities extracted from Wikipedia.
- We apply a machine learning approach to learn the model on a set of user sessions to be subsequently used for queries that are under-represented in historical query logs

Example

- Reformulations rather than completions

- Each potential session starting with q and continued with a particular query reformulations, e.g. $q, q_1, q_{12}, \ldots,$ or $q, q_2, q_{21}, \ldots,$ etc. is a basic mean of *representing a separate aspect or interpretation* of the initial query q.

Problem Goal

- Given the initial query q_0, the goal is to present to the user a *set of suggestions* S_q satisfying the following two conditions:
  - it is *diversified*, i.e., potentially covers many possible interpretations of q_0;
  - *shortens* maximally the subsequent possible sessions to lead the user faster to the satisfactory level of refinement of the query.

Related Work

- Query suggestion:
  - clustering to determine groups of similar queries [Baeza-Yates *et al.*, 2004]
  - entropy models and the use of frequency-inverse query frequency (UF-IQF) [Deng *et al.*, 2009]
  - "Search Shortcuts" [Broccolo *et al.*, 2012]
  - center-piece subgraph that allows for time/space efficient generation of suggestions, also for rare, i.e., long-tail queries [Bonchi *et al.*, 2012]
  - build orthogonal query to satisfy the user's informational need when small perturbations of the original keyword set are insufficient [Vahabi *et al.*, 2013]
- Diversity
  - query refinement is modeled as a stochastic process over the queries [Boldi *et al.*, 2008]
  - diversified query suggestions through pair-wise dissimilarity model between queries [Sydow *et al.*, 2012]
- Machine learning
  - a machine learning approach to learn the probability that a user may find a follow-up query both useful and relevant [Ozertem *et al.*, 2012]

Problem Description

- Given an initial query q, for a subsequent query suggestion q' its *expected shortening utility* can be defined as follows:

$$shortening(q, q') = \sum_{s \in sessions(q,q')} P(s|q) \cdot shortening(s, q')$$

- Lets consider the following options for modeling **P(s|q)** - the likelihood that s will be the subsequent continuation of q:
  - **"cardinality-based likelihood"**:

$$P(s|q) = mult_q(s)/(\sum_{s' \in sessions(q)} mult_q(s'))$$

  - **"weighted likelihood"**:

$$P(s|q) = (len(s) * mult(s))/ \sum_{s' \in sessions(q)} (len(s') * mult_q(s'))$$

  - **"simplistic likelihood"**:

$$P(s|q) = 1$$

Problem Description

- Given an initial query q, for a subsequent query suggestion q' its *expected shortening utility* can be defined as follows:

$$shortening(q, q') = \sum_{s \in sessions(q,q')} P(s|q) \cdot shortening(s, q')$$

- Lets consider the following options for modeling **shortening(s,q')** - the *shortening utility* of suggestion q' for that particular actual continuation s of q:
  - **"absolute shortening"**:

$$shortening(s, q') = pre(s, q')$$

  - **"normalised shortening"**:

$$shortening(s, q') = pre(s, q')/len(s)$$

Problem Generalization

- we define the following set function that models the total shortening achieved by the set of suggestions S_q on all sessions started by q:

$$f(S_q) = \sum_{s \in sessions(q)} P(s|q) \cdot shortening(s, S_q) \quad (1)$$

where

$$shortening(s, S_q) = max_{q' \in S_q} shortening(s, q') \quad (2)$$

- the **MUSETS** problem as an optimization problem:
  - INPUT: Initial, potentially ambiguous query q, number k of suggestions, set C_q of candidate query suggestions for q and a set of recorded sessions $sessions(q)$ that start with q
  - OUTPUT: a $k-$element set S_q of query suggestions that *maximises* the objective function presented in Equation 1.
    Properties: *inherent diversity-awareness, nonfinal queries, non-monotonicity*.
  - It optimizes the *expected* number of steps saved by a user when using suggestions from S_q, in the context of the *unknown* actual interpretation of the ambiguous query q.

Solving the MUSEST Problem

- Standard optimization problem, approached directly by optimizing the objective function
  - the initial query q and sessions started by q are sufficiently represented in query logs
- Machine learning
  - in practice, the sessions starting with q might be insufficiently represented in historical logs
  - this is done in two phases:
    1. Training the model - the training phase we learn the session model with some pre-computed, session-independent representation on queries that are well represented in the historical logs
    2. Evaluation - the second phase, for an incoming query q and some set of *candidate suggestions* C_q we apply the model to predict the shortening utility of each potential suggestion and then construct S_q out of top-k candidate suggestions
  - We are aware that utilizing machine learning model for such a set-wise specification is a challenge, and that our current approach leaves room for improvement that can be tackled in future work.

Machine Learning Approach

- Given a query q', the MUSETS problem aims at predicting a set of query suggestions optimizing a set-wise objective function.
- A challenging task is to represent the queries from a topic point of view.
  - Entity Linking techniques [Ceccarelli *et al.*, 2013].
  - Extended representation of entities from annotated final queries co-occurring in clicked sessions.
- The output space Y is a set of ground-truth labels. We build positive and negative examples as:

$$y_{q'} = \begin{cases} \mathit{shortening}(q, q'), & \text{if } q' \text{ is in a session starting with q;} \\ 0, & \text{otherwise.} \end{cases}$$

- Multiple Additive Regression Trees (MART) [Friedman *et al.*, 2001] optimising Root Mean Squared Error (RMSE).
- The result for each candidate query is a re-ranked list of candidates sorted by decreasing probability of being the suggestion query of the test session.

List of query-related features used to model a *shortening*(q, q').

qi-tokens	The number of tokens in the initial query
qc-tokens	The number of tokens in the candidate query
token-intersection	The intersection of tokens for the two queries
token-union	The union of tokens for the two queries
token-difference1	The difference of tokens between the initial and the candidate query
token-difference2	The difference of tokens between the candidate and the initial query
token-symmetric-difference	The symmetric difference of tokens for the two queries
coocurring-queries-union	The union of co-occurring queries with the initial and the candidate query
cooccuring-queries-intersection	The intersection of co-occurring queries with the initial and the candidate query
difference-qi-qc	The portion of text where the two queries differ, more precisely, the remainder of the candidate query, starting from where it's different from the initial query
qi-substring-of-qc	Reflects whether the initial query is a substring of the candidate query
type-of-query-qc	Reflects whether the candidate query is preponderantly an initial or an inner query
type-of-query-qi	Reflects whether the initial query is preponderantly an initial or an inner query
edit-distance-for-queries	Computes the Levenshtein Distance between the initial and the candidate query
entropy-qi	The entropy of the initial query
entropy-qc	The entropy of the candidate query
probability-qi	The probability of the initial query
probability-qc	The probability of the candidate query
qi-as-qf-probability	The probability of the initial query of being a final query

List of entity-related features used to model a *shortening*(q, q').	
entities-qi	The number of entities found for the initial query
entities-qi-extended	The number of entities for the initial queries computed from annotated co-occurring queries
entities-qc	The number of entities found for the candidate query
entities-union	The union of entities of initial and candidate query
entities-intersection	The union of entities of initial and candidate query
entities-difference1	The difference of entities between the initial query and the candidate query
entities-difference2	The difference of entities between the candidate query and the initial query
entities-symmetric-difference	The symmetric difference of entities between the candidate query and the initial query
entities-union-extended	The union of entities between the extended entity representation of the initial query and the entities of the candidate query
entities-intersection-extended	The intersection of entities between the extended entity representation of the initial query and the entities of the candidate query
entities-difference1-extended	The difference of entities between the extended entity representation of the initial query and the entities of the candidate query
entities-difference2-extended	The difference of entities between the entities of the candidate query and the extended entity representation of the initial query
entities-symmetric-difference-extended	The symmetric difference of entities between the extended entity representation of the initial query and the entities of the candidate query
probability-most-frequent-entity	The probability of the most frequent entity of the initial query in respect to the other entities from the extended entity representation of the initial query
probability-second-most-frequent-entity	The probability of the second most frequent entity of the initial query in respect to the other entities from the extended entity representation of the initial query
probability-third-most-frequent-entity	The probability of the third most frequent entity of the initial query in respect to the other entities from the extended entity representation of the initial query
probability-avg-3-most-frequent-entity	The average probability of the top three most frequent entities of the initial query in respect to the other entities from the extended entity representation of the initial query
entities-with-freq-1	The number of entities with frequency equal to one in the extended entity representation of the initial query

Evaluation

- Data preparation:
  - MSN RFP 2006 query logs
  - Converting all the queries to lowercase, and by removing stop-words and punctuation/control characters
  - Session splitting technique based on the Query Flow Graph
  - Filter out sessions with 3 or less queries
  - Training ($30,000$ sessions) and test set ($2,000$ sessions)
- As a preliminary evaluation, we are reporting below an example of suggestions produced with a MART model learned by using the "simplistic" strategy for modeling $P(s|q)$ and the "absolute shortening" strategy for modeling $shortening(s, q')$.

Results

Query	Candidate Suggestions	$shortening(q, q')$
nemo	finding nemo	0.65
	sea otter	0.07
	great white shark	0.06
	dolphins pictures	0.04
	sea creatures pictures	0.04
	whale sharks	0.03
	nemo pictures	0.03
	finding nemo video clip	0.02
	nemo and friends lamp	0.02
	nemo video	0.02

Table: Example of suggestions derived for the query "nemo" ranked by $shortening(q, q')$.

Results and conclusion

$P(s\|q)$	Metric	Score
Simplistic	NDCG@2	0.7836
	NDCG@5	0.8011
	NDCG@10	0.8214

Table: Results on the test set in terms of NDCG for values of $k \in \{2, 5, 10\}$ for the "Simplistic" strategy.

MUSETS is a promising research direction for modeling shortening of sessions. It is able to produce recommendations that are both relevant and diverse with respect to the query of the user.

Thank you!