1. Description of the data

For our research we used document contents and search engine logs that were generously made available to us by the dutch institute for video and sound [1]. The institute maintains a large database with video fragments with textual description. These video fragments are searched by expert users of the search engine and used in television broadcastings. Users of this service are typically journalists and they pay for this service.

The textual description of the video fragments are the documents that are being searched to retrieve video fragments. They are on average 130 words long, written in Dutch. The documents have three fields, Title, Description and Summary, which are all optional. Documents also have a genre, which can be one of twelve.

2. Collaborative Document Re-Ranking

2.1. **Approach.** When we use collaborative document re-ranking, we will re-rank the documents returned by a search engine based on comparing the user to a number of most similar users and upgrade the ranking score of a document proportional to the user similarity with other users that clicked the documents. As the search engine we are using does not provide any ranking whatsoever, we will assign a ranking sore based on user similarity only. The ranking score for a document d for a given user u_a is defined as:

$$r(d, u_a) = \sum_{u_b \in U} sim(u_a, u_b)\delta(u_b, d)$$

where U denotes the set of k most similar users, $sim(u_a, u_b)$ is the function that computes the user similarity and $\delta(u_b, d) = 1$ if user u_b clicked document d and zero otherwise. A lot of different methods can be used to compute $sim(u_a, u_b)$, we have chosen a language modeling approach in which we define $sim(u_a, u_b)$ as:

$$sim(u_a, u_b) = \frac{1}{|Q_{u_a}|} \sum_{q \in Q_{u_a}} p(q|u_b)$$

where Q_{u_a} is the set of queries from user u_a and $p(q|u)$ can be computed as:

$$p(q|u) = \prod_{q_i \in q} \frac{tf_{q_i,u} + 1}{\sum_{x \in V_u} tf_{x,u}}$$

where V_u denotes the vocabulary of user u. We can define the vocabulary in different ways. We have evaluated two methods of collaborative document re-ranking, the first one uses the Simple User Language Model (SULM) where the user's vocabulary consists only of its queries. In addition to this, we have also defined the vocabulary as the users's queries complemented by all words from the documents the user has clicked. We call this the Extended User Language Model (EULM).

2.2. **Results.** We compare our approach against a random ranking (RANDOM) and greatest hits ranking where documents are ranked according to the number of clicks they got (POP).

2.3. **Discussion.**

2.4. **References.** [1] http://www.beeldengeluid.nl/over-beeld-en-geluid