Background

- Keyword search queries all link to concept(s) that the user may have in mind.
- Each query consists of terms or words, each of which can be related to a concept. E.g., the query “Apple” would only have one term: “Apple”.
- However, the main issue is the ambiguity of a term.
- Many terms can imply different meanings and thus different concepts. For instance, there are two completely different ways to interpret the term “Apple”. One would be linking to the concept [Company], while another would be linking to the concept [Fruit].
- One way to step towards the goal is to determine the distribution of concepts by finding the probability of each concept’s occurring.
- Also, it has been assumed that the distributions of concepts in both a query dataset and a collection are the same. Is that justifiable? Can a collection be a really good approximation?

Goals

- Each term in a keyword query may relate to different concepts in a collection (a compilation of articles, each of which contains concepts).
- In order to effectively answer the keyword queries, we should find the user’s desired concept.
- Thus, given a term, determine its desired concept.

Research Plan

We have selected Wikipedia as a collection for our experiments. We also are using queries from INEX Query Workload (a standard query workload).

- Find the distribution of concepts for articles in Wikipedia.
- Find the distribution of concepts for queries.
- Determine how close the distribution of concepts for queries is to that for Wikipedia.

Research Results

Result:

- Possible rejection of the notion that distributions are the same in both queries and collections.
  - [Country]’s probability was relatively low compared to [Person]’s in the collection, while it was almost equal in the query dataset.
  - [Award]’s and [Conflict]’s probabilities were relatively low in the collection, but relatively high in the query dataset.
  - Suggests that users like to ask about certain concepts more than others, even if they have less variety compared to other concepts.
- May need more sample queries in order to verify that the notion can be rejected.

Related Work/Interaction with Other Projects