Metrics for Assessing Sets of Subtopics

Filip Radlinski
Microsoft Research
Cambridge, UK
filiprad@microsoft.com

Martin Szummer
Microsoft Research
Cambridge, UK
szummer@microsoft.com

Nick Craswell
Microsoft
Redmond, WA, USA
nickcr@microsoft.com

ABSTRACT

To evaluate the diversity of search results, test collections have been developed that identify multiple *intents* for each query. Intents are the different meanings or facets that should be covered in a search results list. This means that topic development involves proposing a set of intents for each query. We propose four measurable properties of query-to-intent mappings, allowing for more principled topic development for such test collections.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Measurement

Keywords: Diversity, Novelty, Subtopic

1. INTRODUCTION

There is increasing interest in producing search results that avoid redundancy, and where results for ambiguous queries cover more of the likely intents for the query (e.g. [3, 4, 5, 6]). A number of researchers have proposed evaluation approaches that take diversity and redundancy into account, including subtopic coverage [1], nugget recall [2] and coverage of Wikipedia disambiguation pages.

However, such evaluation relies on lists of different meanings, aspects or intents for queries. These lists are often created ad-hoc, e.g. by asking human judges to guess possible intents for a query. Given a set of intents generated by some method, there are no well accepted criteria of how the quality of such a set should be measured. We propose four measurable properties for this purpose. We argue that they are desirable for any mapping from queries to intents. We illustrate the properties with a number of examples.

2. FORMALISM

Let q be a query, and i be a textual description of an intent for this query. For example the query "*harry potter*" may have been issued by a user wishing to satisfy the intent "*find the homepage of the official Harry Potter fan club*" or "*buy the first Harry Potter book*". Given a fixed document collection $\mathcal{D}$, let $R_i \subseteq \mathcal{D}$ denote the documents relevant to i in $\mathcal{D}$. Similarly, assume that when a user u issues query q they have a well-defined intent in mind. Let $R_u(q) \subseteq \mathcal{D}$ denote the set of documents relevant to that user.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

Given a query q, suppose some algorithm provides a set $I(q)$ of possible intents for q. Ranking and evaluation is usually done at the per-document level, hence if two intents $i_1, i_2 \in I(q)$ have different textual descriptions but have the same relevant documents, for information retrieval purposes the intents can be considered identical. Thus, we start by defining the similarity between two intents using the Jaccard similarity between their relevant documents:

$$sim(i_1, i_2) = J(R_{i_1}, R_{i_2}) = \frac{|R_{i_1} \cap R_{i_2}|}{|R_{i_1} \cup R_{i_2}|} \quad (1)$$

An alternative similarity measure is Cohen's kappa statistic measured on all judgments for i_1 and i_2. Using the similarity function, we propose four metrics to assess the quality of the mapping $q \rightarrow I(q)$. These properties are that the intents be Coherent, Distinct, Plausible and Complete.

Coherence. Given a query q and intent i, a relevance judge should be able to reliably evaluate the relevance of a document with respect to i. This is desirable because it measures the extent to which the intents produced by the method are themselves unambiguous. This is simply standard interjudge agreement. More formally:

Definition 1. *An intent i for query q is $\alpha-$**coherent** if the documents $R_i^{j_1}$ and $R_i^{j_2}$ judged relevant by two independent judges j_1 and j_2 for intent i satisfy $sim(R_i^{j_1}, R_i^{j_2}) > \alpha$.*

Distinctness. For queries with multiple intents, it is important that the intents be substantially different in terms of the documents judged relevant. Otherwise, judgment effort may be wasted to measure equivalent intents. Formally:

Definition 2. *Intents i_1, i_2 for query q are $\alpha-$**distinct** (for a given document collection $\mathcal{D}$) if $sim(i_1, i_2) \leq 1 - \alpha$.*

Plausibility. To avoid too many intents being provided for a query, all intents provided should be plausible. Intuitively, plausibility means that a sufficient fraction of users who issued q are satisfied with documents that are relevant to i.

Definition 3. *An intent i for query q is $\alpha\beta-$**plausible** if at least fraction α of users who issued the query q satisfy $sim(R_u, R_i) \geq \beta$.*

Note that plausibility could also be defined as: Given an intent i, is q a plausible way of expressing it? This matches how users select queries. However, defining plausibility as the probability of q given R_i would mean that if q is rare, all intents satisfied by popular documents are implausible, as we are not conditioning on the fact that q was issued. Alternatively, estimating it as the probability of q given i makes the

precise wording of i critical, with tiny changes (even ones that don't affect R_i) potentially affecting the plausibility drastically. Our formulation avoids these problems.

Completeness. In generating intents, the frequent meanings of a query should all be present. Otherwise, a diverse ranking for the query may not be useful to many users.

DEFINITION 4. *Given a query q, a set of intents $I(q)$ is considered $\alpha\beta$-**complete** if at least fraction α of users satisfy $\max_{i \in I(q)} sim(R_u, R_i) \geq \beta$.*

Note that the tempting definition, that $I(q)$ is complete if no coherent, plausible and distinct intents can be added, is inappropriate: It would consider $I(q) = \{\}$ complete if q is issued with small probability for many different intents. Instead, our definition requires at least fraction α of users to be satisfied by some intent in $I(q)$.

Also, while plausibility and completeness appear very similar, they are opposites: The former requires that each intent agrees well with the relevance judgments of at least some fraction of users. The later, that for most users there exists at least one intent that agrees well with them.

Since coherence, distinctness and plausibility are defined on individual (or pairs of) intents, we need some way to aggregate given a set $I(q)$. We propose to define each in terms of the minimum of each measure, so that e.g. any implausible intent makes a set of intents implausible.

Finally, while all four metrics are measured on a scale $\alpha, \beta \in [0, 1]$, for any one mapping $q \rightarrow I(q)$ each metric could have different values. Moreover, a "good" value for one metric may be poor for others. We plan to investigate appropriate values for each property in the future.

3. EXAMPLES AND DISCUSSION

We now present four examples of mappings from query to intents (both toy and taken from real datasets). Each illustrates one of the metrics, and how it can be measured.

TOY EXAMPLE 1. **Query:** harry potter
Intents: *(1) Find documents about the movies; (2) Find documents about the books.*

While the intents in this example probably score highly on coherence and distinctness, they would score poorly on **completeness**. Given that many users who issued this query likely have more specific intents, only a small fraction would have their relevance assessment strongly agree with the judgments for these intents. Now consdier adding an intent *(3) Find documents relating to the fictional character*. This new intent is very general, much like the first two. Thus completeness would likely not change. On the other hand, adding *(4) Find showtimes of the latest movie at my local cinema* would improve completeness since some users would likely associate closely with this intent.

Measurability: Completeness of this set could be estimated by asking a some users who issued the query if they agree with document judgments for one of the intents, allowing similarity of R_u with the R_is to be measured.

TOY EXAMPLE 2. **Query***:* bruce croft **Intents***: (1) Find Bruce Croft's homepage; (2) Find a list of Bruce Croft's recent publications; (3) Find a biography about Bruce Croft.*

This example illustrates **distinctness**: While the intents are clearly different from an information need perspective, the set of relevant documents would likely overlap substantially.

Measurability: Given judged documents for each intent, distinctness could easily be measured using Equation (1).

Ambiguous nouns often have disambiguation pages on Wikipedia, which list possible meanings. For example:

REAL EXAMPLE 1. **Query***:* meteor
Intents*: (1) A "shooting star", the visible trace of a meteoroid as it enters the atmosphere; (2) The town of Meteor, Wisconsin, USA; (3) METeOR, Australian information repository; (4) Meteor goldfish, a rare variety of goldfish; ... (8) Ireland's third mobile phone operator; (9) Meteor Vineyard in Napa Valley, USA; ... (11) A series of weather satellites of the Soviet Union; ... (27) A 1979 science fiction film; ... (39) BSA Meteor Air Rifle.*

While the structure of Wikipedia means the intents are likely coherent, distinct and complete, we now consider **plausibility**. Intuitively, the list of intents is long, thus perhaps some are implausible – but this is not necessarily the case.

Measurability: By asking a sample of users who issued a particular query (or asking study participants to imagine issuing it) to specify if they agree with particular intents, plausibility of each of a set of intents could be assessed.

REAL EXAMPLE 2. **Query***:* appraisals **Intents***: (1) What companies can give an appraisal of my home's value? (2) I'm looking for companies that appraise jewelry; (3) Find examples of employee performance appraisals; (4) I'm looking for web sites that do antique appraisals.*

This example is TREC topic number 8 from the 2009 Web Track, and illustrates **coherence**.

Measurability: Inter-judge agreement can be measured by obtaining document judgments for each intent from two judges. Note that the intents may differ in agreement rate.

4. CONCLUSION

We have proposed four properties of a mapping from query to a set of intents. They can be used as guiding principles in developing evaluation sets that take account of query ambiguity and diversity, and are measurable given a collection judged for different intents – although also working with actual users appears essential to measure some of the metrics.

5. REFERENCES

[1] C. Zhai, W. Cohen, and J. Lafferty. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In *SIGIR*, 2003.

[2] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. In *SIGIR*, 2008.

[3] H. Chen and D. Karger. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. In *SIGIR*, 2006.

[4] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *SIGIR*, 1998.

[5] F. Radlinski, P. Bennett, B. Carterette, and T. Joachims. SIGIR Workshop Report: Redundancy, Diversity and Interdependent Document Relevance. *SIGIR Forum*, 43(2):46–52, 2009.

[6] Y. Yue and T. Joachims. Predicting Diverse Subsets using Structural SVMs. In *ICML*, 2008.