# Temporal Variance of Intents in Multi-faceted Event-driven Information Needs*

Stewart Whiting, Ke Zhou & Joemon M. Jose
School of Computing Science
University of Glasgow, UK
{stewh,zhouke,jj}@dcs.gla.ac.uk

Mounia Lalmas
Yahoo! Labs
Barcelona, Spain
mounia@acm.org

## ABSTRACT

Time is often important for understanding user intent during search activity, especially for information needs related to event-driven topics. Diversity for multi-faceted information needs ensures that ranked documents optimally cover multiple facets when a user's intent is uncertain. Effective diversity is reliant on methods to (i) discover and represent facets, and (ii) determine how likely each facet is the user's intent (i.e., its popularity). Past work has developed several techniques addressing these issues, however, they have concentrated on static approaches which do not consider the temporal nature of new and evolving intents and their popularity. In many cases, what a user expects may change dramatically over time as events develop. In this work we study the temporal variance of search intents for event-driven information needs using Wikipedia. First, we model intents based upon the structure represented by the section hierarchy of Wikipedia articles closely related to the information need. Using this technique, we investigate whether temporal changes in the content structure, i.e. in a section's text, reflect the temporal popularity of the intent. We map intents taken from a query-log (as ground-truth) to Wikipedia article sections and found that a large proportion are indeed reflected in topic-related article structure. By correlating the change activity of each section with the use of the intent query over time, we found that section change activity does reflect temporal popularity of many intents. Furthermore, we show that popularity between intents changes over time for event-driven topics.

**Categories and Subject Descriptors:** H.3.3 [Information Search and Retrieval - Search Process, Selection Process]
**Keywords:** Intent; Diversity; Temporal; Time; Events

## 1. INTRODUCTION

Under-specified or ambiguous queries are a common problem for web information retrieval systems [2], especially when the queries used are often only a few words in length. In some cases the user may provide an information need which is *ambiguous* as it has many interpretations. For instance, the query 'jaguar' could refer to the car, animal or operating system. On the other hand, the information need may have multiple *facets* (or synonymously, search *intents*), each covering a specific part of the broader topic. The user's actual intent may be one or more of these facets. For instance, the query 'steve jobs' could refer to facets such as his death, biography, movie or keynotes, etc. Without further clarification it is impossible to know the actual intent of the user. To alleviate this problem, result diversification is a common strategy employed to maximise average retrieval effectiveness for all users. For multi-faceted queries, diversity attempts to optimally rank documents to cover as many possible search intents of the query in the least results possible.

Time plays a central role in information retrieval (IR) as information needs may often arise from recent media interaction, stimulating users to seek information about topics related to ongoing events and phenomena [1]. Many queries have a temporal affinity affecting popularity and in many cases, intent over time [5, 9]. Previous work has demonstrated the temporal variance in popularity of ambiguous search intents [9]. In this work we focus on studying the temporal variance of search intents in multi-faceted information needs. This poses two challenges: (i) discovering and representing multi-faceted search intents that evolve over time, and (ii) the likelihood that each search intent will be that which the user is interested in over time.

Despite ongoing temporal change, existing approaches to multi-faceted intent diversity assume a static set of search intents and likelihood (popularity). Particularly for event-driven topics which are ever changing, this may lead to a sub-optimal ranking. For example, consider the 2011 Thailand Floods (illustrated in Figure 1). At first, people were interested in finding out about the local impact. As the media reported the impact of flooding on manufacturing facilities, attention shifted to the impact on hard drive supply and prices. As the supply issues eased, focus was once again on dealing with the aftermath and how to avoid problems again.

While event-driven topics are occurring, there will be either very little or no past evidence available for search intents and likelihood; and old evidence may not reflect real-time user expectations. Developing IR systems which can respond in real-time to changing user expectations motivates the need to find reliable sources of recent event information. For this reason, we look towards a large-scale real-time collaboratively edited encyclopaedia: Wikipedia[1], as a means of modelling and representing multi-faceted intents and their temporal variance. With users collaborating globally, major events and phenomena are often described shortly after their occurrence [7, 6]. Previous work has exploited Wikipedia taxonomy structure and page linking for representing non-temporal intents for diversity [4]. However, to model and represent finer-grained in-

*This research is partially supported by the EU-funded project: LiMoSINe (288024).


*SIGIR'13,* July 28–August 1, 2013, Dublin, Ireland.


[1] http://www.wikipedia.org

tents of event-driven information needs, we propose to exploit the structure of sections contained in articles related to the information need. Furthermore, based on the conjecture that sections which change frequently are popular, we posit that the change activity of a section reflects its likelihood as a search intent. Henceforth, we motivate two research questions for this paper:

**RQ1**. Does Wikipedia article structure reflect the search intents for multi-faceted queries (i.e. Do sections/subsections each reflect an intent)?

**RQ2**. Does the evolution of an article (i.e., sections being created, changed, etc.) reflect intent popularity over time?

Our contribution is three-fold: to the best of our knowledge, this is a first attempt to (i) investigate diversity for event-driven queries, (ii) use the stream of Wikipedia article changes to investigate temporal intent variance for event-driven queries[2], and (iii) quantify temporal variance between a set of search intents for a topic.

# 2. RELATED WORK

Recent IR research has investigated "diversity-based", or, "subtopic" retrieval approaches for modelling user search intents during search tasks [3, 8], where ambiguity or multi-faceted information needs cause relevance uncertainty. An intent-diversified result ranking can be created by interleaving documents sampled from possible search intents, with the importance of each intent indicated by several features such as prior search intent click-through rate or original document relevance. Although the temporal variance of multi-faceted intents and their popularity has been acknowledged [5], little work has quantitatively studied the temporal variability of multi-faceted intents, assuming they maintain static over time.

The implications of time and temporality has been studied in many IR problems. Adar at al. [1] quantitatively studied the temporal correlation of topics appearing across different media (e.g. blogs, news and television listings) and search engines as interest in ongoing events and phenomena spreads. Kulkarni et al. [5] conclude that many web search queries and relevant documents are influenced by a periodic or real-time event-driven temporal dimension. Repeatedly sourced relevance judgements suggest temporally-sensitive information needs and query intent, as a consequence of events. In past work [9], we studied the temporal variability inherent in many ambiguous information needs, and simulated the effect on diversity evaluation. In this work, we aim to observe how multi-faceted search intents also vary over time.

Wikipedia article editing activity and page view streams have seen interest in recent event detection and tracking work [6, 7]. Hu at al. [4] utilised Wikipedia taxonomy, article link structure and entire articles to statically represent the most likely intents for any information need.

# 3. RQ1: REPRESENTING INTENTS

To address our research questions, we must first represent multi-faceted search intents so that we can evaluate Wikipedia-derived intents (i.e. those appearing in Wikipedia article section structure) against ground-truth search intents.

As we are studying events following their occurrence, for simplicity at this stage we assume that all ground-truth search intents are present during the event. Later, in Section 4 we investigate search intent popularity over time.

We select a set of 20 event-driven queries/topics to test our hypothesis, outlined in Section 3.1. In Section 3.2 we describe how to obtain the ground-truth of possible search intents. Further, in Section 3.3 we describe our methodology for assessing the matches between ground-truth search intents and Wikipedia article sections. Finally, in Section 3.4 we present results of representing search intents using Wikipedia sections.

## 3.1 Queries

We select two categories of event-driven queries for testing our hypothesis, related to significant events between January 2010 and December 2012. 4 individuals were asked to identify major events in the Wikipedia 2011-2012 News pages[3], and provide the query they would use to find general information on the event. From this pool of events and queries we selected two sets of 10 topics based on the following characteristics.

All topics are themselves an event, with each having a central descriptive Wikipedia article. Topics 1-10 are relatively short events (e.g., severe weather or a shooting), which have most temporal interest between 1 to 14 days. In contrast, topics 11-20 are prolonged events which happen over many weeks, months or even years (e.g., the Libya Intervention). Often these longer events are composed of many facets, concerning different people, places and interaction over time[4]. Two example categorised queries with their multi-faceted intents are presented in Table 1. The underlying reason to choose these two categories of queries is to reflect events with different temporal characteristics, particularly for investigating RQ2 in Section 4.

Table 1: Example short- and long-term event-driven queries, along with their multiple facets (obtained from Google Related Searches).

| Query (topic) | Facets (from query-log) |
|---|---|
| Eyjafjallajokull *Short-term (Topics 1-10)* | eyjafjallajokull **effects**, eyjafjallajokull **facts**, eyjafjallajokull **volcano webcam**, how to **pronounce** eyjafjallajokull, eyjafjallajokull **bbc**, eyjafjallajokull **case study** |
| Libya Intervention *Long-term (Topics 11-20)* | libya intervention **responsibility to protect**, libya intervention **poll**, libya intervention **debate**, libya intervention **timeline**, libya intervention **nato**, libya intervention **legality**, libya intervention **oil**, libya intervention **success** |

## 3.2 Search Intents

We first obtain the ground-truth of search intents for each event-driven query. For large-scale commercial search engines, the ground-truth of intents should be based on a large number of users. Since we do not have a query-log, we instead propose an approach to derive intent ground-truth using features provided by a commercial search engine, i.e. Google. We examined the suggestions provided by Google Query Auto-Completion, Google Related Searches and Google Trends Related Searches. To select the best source, we define the criteria as follows: (i) the source should cover a variety of diverse intents/facets of an event, and (ii) it should cover the most popular intents/facets so that temporal statistics can be obtained. Based on our observation, we believe that the queries suggested by Google Related Search met our criteria and therefore we chose this as the ground-truth. Google Auto-Completion and Google Trends Related Searches data either over-reward tail queries or do not cover multiple diverse facets. An example of ground-truth facets obtained in this way is shown in Table 1.

## 3.3 Intent Matching and Assessments

To establish which ground-truth search intents are reflected by Wikipedia article sections, we attempt to match each ground-truth

[2]Note that the methodology used within this paper can also be applied to general, non-event driven queries. We leave broader query categories to future work.

[3]http://en.wikipedia.org/wiki/[2010|2011|2012]

[4]In this work we do not investigate seasonal event-driven queries such as Christmas. We are interested in events that are harder to predict given a lack of past evidence.

intent to a possible section and furthermore, assess the match strength. This consists of several steps: (i) *Event Article Identification*: identifying multiple Wikipedia articles that are most related to event-driven topics, (ii) *Section-Intent Automatic Matching*: retrieving sections from the articles identified above, that might match the intents (for further assessments), and (iii) *Match Assessments*: assessing match strength between retrieved sections and search intents. We illustrate each step in detail as below.

**Event Article Identification.** Before listing all the candidate sections that can be potentially matched to the search intents, the set of Wikipedia articles most related to each event-driven topic, $\{A_{topic}\}$, must be identified. Major events are typically represented by a central article (e.g. 'Occupy Movement'), with related articles detailing substantial aspects such as 'Reactions to the Occupy Movement', 'Occupy Movement in the United States' and 'Occupy Canada'. As this work concentrates on a small number of topics we manually identified related articles as those linked from the central article via 'See also:' and 'Main article:' references, although past work has proposed automatic methods [4].

**Section-Intent Auto-Matching.** We posit that a search intent is reflected by one or more sections (or, subsections) contained in $\{A_{topic}\}$. For example, the 'Occupy Movement' article has sections including 'background', 'we are the 99%', 'goals', 'methods', and 'protests' (with a subsection for each participating country). Despite the hierarchical nested structure of Wikipedia sections, to avoid complexity we employ a flat section structure. Hierarchy is particularly challenging for Wikipedia articles as it will change dramatically over time. As such, we leave this issue open for further work. Matching was performed semi-automatically. To begin, we took each ground-truth search intent and extracted the intent key terms and automatically retrieved up to three sections from $\{A_{topic}\}$ which most contained the term in their header title or text. For example, for the search intent 'libya intervention oil', we identified sections referring to the word 'oil'.

**Match Assessments.** With the large pool of potentially matched sections retrieved by the system described above, two separate individuals were asked to annotate the extent to which each section reflected the intent. Assessments were made in three grades: either a *hard* match (i.e., section is entirely about the intent), *soft* match (i.e., loosely related) or *no* match. To prepare the final ground-truth intent and Wikipedia section matches, annotation conflicts were resolved by choosing the lower of the two labels (e.g., a 'strong' and 'weak' were resolved to 'weak').

## 3.4 Evaluation

To evaluate the effectiveness of search intent representation in Wikipedia article sections against the ground-truth, we report $Recall_{wiki}$ (i.e., # intents that Wikipedia covers / # total intents). It is calculated per-topic, however we report the average.

In Table 2, we can observe that Wikipedia sections substantially reflect user's search intents as it has high $Recall_{wiki}$ for both hard (0.68) and soft (0.87) assessments, which answers our **RQ1**. From close examination, the search intents without coverage are generally intents that are related to a specific resource (e.g. "bbc"), or generic type of information (e.g. "jokes"), and so are missing from Wikipedia. These search intents are less likely to have a significant temporal variance as they refer to generic facets common to many event-driven topics.

# 4. RQ2: TEMPORAL INTENT VARIANCE

In the previous section we evaluated representing multi-faceted search intents with Wikipedia article structure. In this section, we observe how the temporal variance in search intent popularity is reflected by section change activity in Wikipedia. In Sections 4.1 and 4.2, we compare the ground-truth temporal popularity of a search intent (obtained from the query-log) to the temporal volume of changes made to the intent representation in Wikipedia, i.e., the frequency of changes made to the text content in the article section. Furthermore, we quantify the extent to which all search intents for a topic vary in popularity over time.

Table 2: $Recall_{wiki}$ of soft and hard matched intents, for topics 1-20 (All), 1-10, and 11-20.

| | Soft | Hard |
|---|---|---|
| Topics | $Recall_{wiki}$ | $Recall_{wiki}$ |
| *All* | 0.89 | 0.68 |
| *1-10* | 0.87 | 0.64 |
| *11-20* | 0.92 | 0.72 |

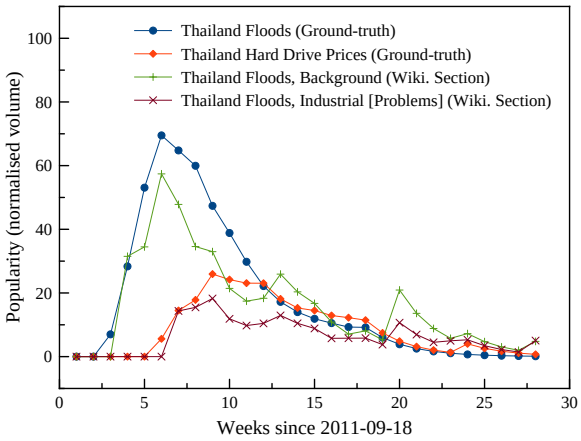


Figure 1: Temporal popularity from 2011-09-18 for two search intents of the Thailand Floods. Each time-series was normalised (maintaining magnitude) and exponentially smoothed ($\alpha = 0.35$).

## 4.1 Experiment

Comparing the temporal variance of ground-truth search intents and Wikipedia article sections requires comparable time-series of query volume and section change activity.

**Ground-truth.** Query-logs capture the popularity of facets over time [5]. As we do not have our own large-scale query-log we rely on the temporal query volume data provided by Google Trends[5] as the ground-truth temporal popularity for each ground-truth search intent query. An example of this is shown in Figure 1.

**In Wikipedia.** Temporal changes of Wikipedia article sections can be obtained by comparing contiguous Wikipedia article revisions. The stream of changed sections can be mined from the stream of article revision text. Standard `diff` and `patch` operations identify locations of text changes between adjacent revisions. Each change can in turn be resolved to a specific section by seeking its nearest parent section title header.

Section structure (e.g. section presence and hierarchy) is constantly evolving during collaborative editing. In some cases, this poses challenges for identifying the relevant section for a text change. We leave the issues raised by this to future work.

**Evaluation.** To compare the ground-truth intent query popularity and Wikipedia article section change activity, we aggregate the noisy continuous stream of data points using temporal buckets to group each time-series into a lower-dimensional series. Each new data point represents an $N$ day period (where $N = 1, 7, 14$ days).

[5] `http://www.google.com/trends`

Pearson's correlation co-efficient, $r$, is used to measure the temporal similarity between the popularity time-series.

We quantify the variance between search intent popularity over time as follows. For each topic's search intents, with a temporal bucket size of $N$ days ($N = 1, 2, 7$) we rank sections mapped to intents by their section change frequency (i.e., popularity) over each bucket period. Spearman's rank correlation $\rho$ is then computed between adjacent bucket ranks, providing a measure of period to period intent popularity similarity. Rank-based measurement is relatively robust to noise, and background fluctuations (e.g. weekends). Periods of intent popularity instability are reflected by a low Spearman's $\rho$, conversely, periods of stable intent popularity are reflected by $\rho \approx 1$. For each topic, we aggregate the average Spearman's $\rho$ from all adjacent buckets over time, discarding those where $\leq 1$ sections were changed. We treat any $\rho < 0$ as $\rho = 0$.

## 4.2 Results

In Table 3 we report the average correlation $r$ between the ground-truth and Wikipedia intent representation popularity, for each temporal bucket size, and topics 1-10 and 11-20.

Considering the raw Wikipedia article change stream is relatively noisy (e.g. if one editor repeatedly commits tiny changes), short-term topics have a relatively strong correlation at all temporal bucket sizes. As the bucket size increases, correlation is increased as daily noise is aggregated and smoothed. For long-term events, correlation increases with a larger bucket size, e.g. 14 days. This is likely caused not only by noise smoothing, but also the fact that longer events may consist of aspects which develop over many weeks rather than just days.

Table 3: Average temporal correlation between Wikipedia section change activity and ground-truth query popularity.

| | Average Pearson $r$ | | |
|---|---|---|---|
| | Temporal bucket size | | |
| Topics | 1 day | 7 days | 14 days |
| *1-10* | 0.32 | 0.49 | 0.58 |
| *11-20* | 0.15 | 0.25 | 0.33 |

Table 4: Temporal variance of topic intent popularity, indicated by mean average Spearman's $\rho$ (inc. stdev.), for different bucket sizes.

| | Mean Average Spearman's $\rho$ for Topics | | |
|---|---|---|---|
| | Temporal bucket size | | |
| Topics | 1 day | 2 days | 7 days |
| *1-10* | 0.08 ($\pm$ 0.12) | 0.07 ($\pm$ 0.14) | 0.12 ($\pm$ 0.16) |
| *11-20* | 0.03 ($\pm$ 0.07) | 0.09 ($\pm$ 0.09) | 0.2 ($\pm$ 0.2) |

**Quantifying Temporal Variance.** We report the mean of the average Spearman's $\rho$ (MA$\rho$) for each temporal bucket period, topics 1-10 and 11-20 in Table 4. For both short- and long-term events, intent popularity varies substantially over time, indicated by a relatively low MA$\rho$ for all bucket size periods. However, the relatively large dispersion (denoted by $\pm$) suggests temporal variance differs considerably between topics. Overall low MA$\rho$ and per-topic inconsistency may to some extent be caused by articles which lack a large volume of changes during certain periods, introducing noise especially at shorter bucket periods (e.g. 1-2 days).

Long-term topic intents (i.e. topics 11-20) vary considerably less over 7 day periods compared to 1-2 days. This may be due to a few intents becoming established and maintaining popularity over longer periods, compared to the noise present at shorter bucket periods caused by non-event related section changes [7]. In comparison, short-term events (i.e. topics 1-10) have marginally less intent popularity variance during short periods, as the event quickly and consistently develops. However, there is little increase in MA$\rho$ observed at longer bucket periods as there are so few changes made outside the short duration of the event.

# 5. DISCUSSION & CONCLUSION

We have addressed the research questions outlined in Section 1 through preliminary experiments and analysis. We mined multi-faceted ground-truth search intents from the Google Related Searches and matched them to sections contained in related Wikipedia articles. Furthermore, we studied ground-truth temporal correlation, and quantified temporal intent variance.

In **RQ1** we hypothesised that Wikipedia article sections can be used to represent search intents. Results presented in Table 2 demonstrate that there is substantial overlap of intents, suggested by a high recall. Although not all search intents are reflected by Wikipedia (e.g. 'jokes'), major informational intents are usually present as one or more article sections.

Extending **RQ1**, the aim of **RQ2** was to study whether search intent popularity is reflected by article section change activity, i.e., popular sections/intents are changed more frequently. Results presented in Table 3 suggest a medium to strong correlation for both short- and long-term events. Short-term events correlate daily with search intent popularity, whereas long-term events only correlate over longer temporal intervals (e.g. 14 day periods), echoing their slower but continued development over longer periods of time. Furthermore, Table 4 illustrates the daily and weekly variance of intent popularity. In particular, the popularity of short-term event intents changes quickly, whereas long-term events have some intents which remain more stable over longer periods.

**Conclusion.** In this work, we have studied how search intents can be represented by Wikipedia article sections. We have established that many search intents for event-driven topics (e.g. news events) can be represented by Wikipedia article sections. Moreover, the popularity of each of these intents varies over time and is reflected by the editing activity of each section. With little or no query-log evidence, Wikipedia article structure offers a means to understand (i) the search intents currently present and emerging, and (ii) the temporal popularity of each intent.

Temporal intent variance motivates time-aware IR model development. There is much scope to improve this model of representing search intents and popularity using evolving Wikipedia structure.

# 6. REFERENCES


[1] E. Adar, D. S. Weld, B. N. Bershad, and S. S. Gribble. Why we search: visualizing and predicting user behavior. In *WWW '07*, pages 161–170, New York, NY, USA, 2007. ACM.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09*, pages 5–14, New York, NY, USA, 2009. ACM.

[3] J. G. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98*, pages 335–336, 1998.

[4] J. Hu, G. Wang, F. Lochovsky, J. tao Sun, and Z. Chen. Understanding user's query intent with wikipedia. In *WWW '09*, pages 471–480, New York, NY, USA, 2009. ACM.

[5] A. Kulkarni, J. Teevan, K. M. Svore, and S. T. Dumais. Understanding temporal query dynamics. In *ACM WSDM '11*, pages 167–176, New York, NY, USA, 2011. ACM.

[6] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First story detection using twitter and wikipedia. In *SIGIR '12*, Time-aware Information Access Workshop '12, 2012.

[7] S. Whiting, K. Zhou, J. M. Jose, O. Alonso, and T. Leelanupab. Crowdtiles: presenting crowd-based information for event-driven information needs. In *CIKM '12*, pages 2698–2700, 2012.

[8] C. Zhai, W. W. Cohen, and J. D. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03*, pages 10–17, 2003.

[9] K. Zhou, S. Whiting, J. M. Jose, and M. Lalmas. The impact of temporal intent variability on diversity evaluation. In *ECIR '13*, pages 820–823, Berlin, Heidelberg, 2013. Springer-Verlag.