Automatic Generation of Topic Pages using Query-based Aspect Models

Niranjan Balasubramanian
University of Massachusetts Amherst
140 Governors Drive
Amherst, MA 01003
niranjan@cs.umass.edu

Silviu Cucerzan
Microsoft Research
1 Microsoft Way
Redmond, WA 98052
silviu@microsoft.com

ABSTRACT

We investigate the automatic generation of *topic pages* as an alternative to the current Web search paradigm. We describe a general framework, which combines query log analysis to build aspect models, sentence selection methods for identifying relevant and non-redundant Web sentences, and a technique for sentence ordering. We evaluate our approach on biographical topics both automatically and manually, by using Wikipedia as reference.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Measurement.

Keywords: topic pages, aspect models, query logs.

1. INTRODUCTION

We define topic pages as content hubs, which aggregate *useful information* on particular topics from sources all over the Web, and enable users to easily navigate to those sources. They collect and organize information pertaining to different aspects of a topic in one place, explicitly addressing redundancy and diversity in addition to relevance. Figure 1 shows an example topic page automatically generated by our system for the topic "William Shatner", which covers distinct relevant aspects, such as his acting career, famous movies, books, and TV commercials, and provides links to various Web sources for additional information on each aspect. In contrast to typical Web search result pages, whose snippets often provide superficial and/or redundant information, topic pages present a single high-quality summary while retaining pointers to a multitude of information sources.

One of the key challenges in generating topic pages is defining the notion of information usefulness. Wikipedia's content harvests the wisdom of the crowds, but with a bias toward what the active contributors believe is important for a topic. Multi-document summarization systems often define the information usefulness of sentences based on properties of the summarized documents themselves. In contrast, we employ Web search query logs to build *aspect models* that capture a consensus of user interests with respect to the topics. We then attempt to cover these aspects accurately and non-redundantly, task which includes sentence retrieval and ranking, as well as coherent aggregation of the information.

CIKM'09, November 2–6, 2009, Hong Kong, China.

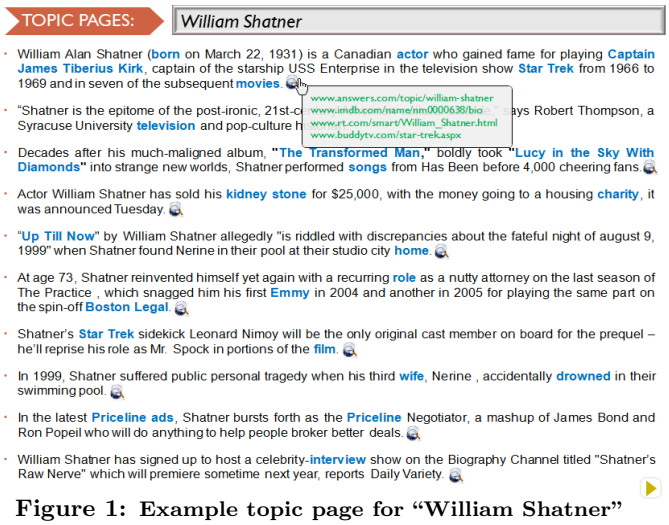


Figure 1: Example topic page for "William Shatner"

2. RELATED WORK

Topic page generation could be viewed as a topic-focused **multi-document summarization** task. However, a crucial difference is that typical multi-document summarization systems utilize a pool of given text documents and employ word-based statistics from the input pool directly to determine sentences that need to be added to the summary [8]. While the documents obtained through Web search by querying a topic are often related to that topic, they frequently contain irrelevant information and lack cohesiveness, which makes them difficult to summarize [7]. Instead of using the documents alone, Lacatusu et al.'s system [6] breaks down a complex query into simpler queries and produces summaries as responses to those queries. In a similar fashion, we generate aspect models for a topic from query logs, and employ them to construct the topic pages, thereby avoiding the difficult task of summarizing non-cohesive Web documents.

Early work on **biography generation** has focussed on multi-document summarization of information from news collections [10, 11]. Alani et al. [1] use pre-defined biographical templates to collect biographical facts, Filatova and Prager [5] identify person-specific, occupation-specific, and general biographical events, while Biadsy et al. [3] learn a biographical sentence classifier from Wikipedia and TDT4. Conversely, we extract from query logs aspect models with varying degrees of specificity to the target topic and similar topics, without the need to explicitly capture occupational roles, templates, or contextual patterns.

In contrast to **Web search results clustering** (e.g., [12]), which implicitly attempts to discover sub-topics within search results, we use explicit aspects derived from query logs to retrieve relevant sentences and then organize them in a co-

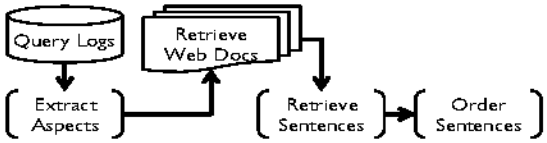


Figure 2: System architecture for topic page generation

herent fashion, with pointers to their Web sources. We opted for aspect models with low complexity, but more sophisticated models, such as approaches for class attribute extraction and propagation in conceptual hierarchies [9], could also be employed in our topic page generation framework.

3. OVERVIEW

We built a reference collection from Wikipedia to develop, train, and evaluate the proposed system and its subcomponents. To select biographical topics, we first gathered the Wikipedia pages that were labeled with the category "Living people". We used the page titles, after removing parentheticals, as topic names. Less than 5% of these names were duplicates, indicating possible ambiguity problems, and were eliminated from the candidate set to simplify the experimental setup.[1] Because entertainment-related topics appeared to dominate this set, we inspected the category information for a small set of known topics to identify categories corresponding to diverse occupations. For example, to build a seed list of politicians, we selected pages in the "Living people" category with additional category labels such as "U.S. Senator". Then, we propagated the occupation labels to other topics by using the Wikipedia category sets. Additionally, to avoid topics with very little user interest, we removed topics that had fewer than 20 entries in a six-month query log from the Bing search engine. From the resulting topics, we randomly selected uniformly over occupation labels three disjoint sets of 100 topics for training, development and test.

Figure 2 shows the key components of our general framework for generating topic pages: aspect extraction, content retrieval/selection, and content organization.

4. ASPECT MODELS

Web search query logs can be seen as aggregators of the information needs of a vast number of users with respect to any topic. However, the distribution of queried aspects can be heavily biased towards events occurring within the time frame of the analyzed logs. Additionally, similar information needs can be expressed using multiple lexical choices. As a result, many highly popular terms co-occurring with a topic in user queries refer to the same aspect. Our experiments showed that term clustering techniques do not address these problems in a satisfactory manner. Thus, we examined an approach that employs three types of query-log-derived aspects: *self* (specific to the topic), *related* (common across related topics), and *general* (common to all topics). To build the *self aspect model* for a topic, we first extract queries that contain that topic. Then, we select the most frequent n terms that occur in those queries after filtering out stopwords.[2] To generate the *related aspect model* for a topic, we sort all topics in our pool based on the similarity of their individual self aspect models to the self aspect model of the given topic. We then combine the self aspect models of top m ranked topics and select the top n terms from the combined model. Finally, we build a *general aspect model* by combining the self aspect models of all the topics in our pool.[3] Since this model is generated only once, we manually filter out biographically-meaningless terms such as "official page" before selecting the top n of the remaining terms.

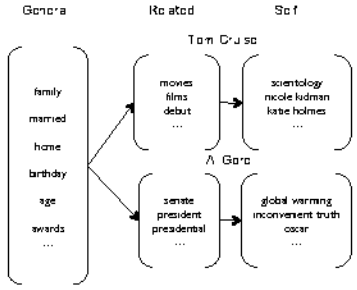


Figure 3: Aspect models for "Tom Cruise", "Al Gore"

Figure 3 exemplifies these query-log-derived aspect models for the topics "Tom Cruise" and "Al Gore".

To empirically determine suitable assignments for parameters m and n, we tried a range of values on the development set and compared the self and related aspect models with the concepts occurring in Wikipedia pages. To perform the comparison, we employed the title pages of Wikipedia as our concept space and the anchor text of the Wikipedia links as the vocabulary for these concepts. The related aspects have both higher precision and higher recall of Wikipedia concepts than the self aspects. Increasing the number of aspects in the models from 10 to 30 produces recall increases from 4% to 8% for the self model and from 6% to 10% for related aspects, for a relatively smaller precision trade-off (from 31% to 27% for self and from 38% to 28% for related).

5. SENTENCE SELECTION

We now address the task of retrieving sentences from the Web for the aspects pertaining to a topic, and focus on grammaticallity, relevance, non-redundancy and diversity.

5.1 Sentence Grammaticallity

To extract sentences from Web documents, we use a simple html parser and a sentence boundary detector based on regular expressions and word casing statistics. This process often results in extracting ungrammatical sentences due to html misparsing, failure of the boundary detector to handle html/script content, or ungrammatical content such as blog postings and user responses to well written articles. To correct this, we used lexical indicators, language modeling, and perplexity features (as displayed in Table 1) to train a logistic regression classifier, which achieved more than 80% precision at 85% recall in identifying grammatical sentences.

Table 1: Junk Classifier Features

Perplexity	Alphabet to special chars ratio
Unigram Likelihood	Alphabet to numbers ratio
Bigram Likelihood	Caps to Lower case ratio
Contains URLs	Link patterns (- - -, | | |, . . .)

5.2 Sentence Relevance

For an aspect vector, given a pool of candidate sentences, we have to select a subset that covers the aspects of interest, preferably one sentence per aspect. An approach that ranks all candidate sentences based on their similarity with the entire aspect vector (the *Full Aspect* method) is prone to selecting long and highly redundant sentences, which mention multiple aspects. A competing straight-forward approach is to select for each individual aspect one sentence that contains both the topic and the aspect. However, in addition to

[1] Preliminary experiments show this type of ambiguity can be handled by systems such as [4] to disambiguate the targeted name for each Web page retrieved.

[2] To capture not only words but also concepts, we employ for query tokenization a list of Wikipedia concepts. Thus, our models contain both words and phrases.

[3] Because we target only biographical topics, we can assume that they share a common outline and there is no need for extra topic pre-categorization.

Table 2: Example sentences for the topic "Adrien Brody" and the aspect "The Pianist".

Adrien Brody, best known for his Oscar-winning performance in "**The Pianist**," was in the audience Friday at the opening.
Rachel Weisz (The Constant Gardener), **Adrien Brody** (**The Pianist**), Mark Ruffalo (Zodiac), and Rinko Kikuchi (Babel) star in THE BROTHERS BlOOM [...]
Adrien Brody received widespread recognition when he was cast as the lead in Roman Polanski's **The Pianist** (2002).
Adrien Brody is a New York actor who is known to international audiences as the star of Roman Polanski's 2002 film, **The Pianist**.

overlooking sentences with only pronominal references, this approach cannot ensure that the focus of the selected sentence is the targeted aspect. The mere presence of the topic and the aspect is not always an adequate indicator of relevance. For example, Table 2 shows four sentences extracted for the topic "Adrien Brody" and the aspect "The Pianist", of which the first two do not focus on the targeted aspect.

To identify sentences that focus on the connection between the topic and an aspect, we make use of *aspect-specific contexts*, which rely on the observation that the contexts in which an aspect occurs often differ from the contexts of other aspects, as well as the overall topic context. Therefore, we first build an aspect-specific context vector for each aspect by extracting the terms in all sentences containing the topic and the aspect. These give more weight to words related to the aspect than to noisy words from sentences that have other foci (e.g., the context vector for "Adrien Brody" and "The Pianist" contains entries for "Oscar", "Roman Polanski", "won", "award", "star", etc.) Then, we interpolate the aspect-specific context vector with the entire aspect vector. Finally, we compute the vectorial similarity of the candidate sentences with the corresponding context vector.

5.3 Redundancy and Diversity

An important feature of the proposed topic pages is the non-redundant coverage of aspects. Redundancy is caused mainly by the aspect models extracted from query logs and by sentences that cover multiple aspects. Using the relevance scores alone may lead to selecting sentences that share aspects and common vocabulary. To remove redundancy and promote selection of novel sentences, we can adopt the techniques used in novelty detection work [2], and gather iteratively a set of sentences by adding to the set a new sentence based on a linear combination of its relevance score and its novelty score with respect to the current set (the *Novelty* method). However, direct application of such a technique does not ensure the coverage of all the aspects of interest.

To enforce both diversity and novelty in the sentence selection process, we also investigated the following methods:
Typical: For each aspect, extract sentences that contain the aspect and the topic, and build a context vector. Re-rank all sentences based on their cosine similarity to a linear interpolation of the aspect vector and the new context vector.
Diversity: Starting with the full aspect vector, iteratively select one sentence that is most similar to the aspect vector. After each selection, modify the aspect vector by down-weighting the aspects covered in the selected sentence by δ. Repeat until the desired number of sentences are selected.
Diversity+Typical (D-T): Start with the full aspect vector. Select an aspect from the full aspect vector. Use the Typical method to get the best candidate sentence for the aspect. Then, remove all the aspects covered by the selected sentence. Repeat the process until no more aspects remain in the vector or the desired number of sentences are selected.

6. EXPERIMENTS

Given a set of topics and their corresponding aspect models, we conduct sentence selection experiments by employing Wikipedia sentences.[4] We use both term-based and sentence-level metrics. To favor diversity, we also employ modified *D-metrics*, for which we allow each reference sentence to be matched by at most one of the selected sentences.

6.1 Aspect Models Combination

During training we allow each method to learn its own weights for interpolating the self (S), related (R) and general (G) aspect models: $A_m = \beta S + \gamma R + (1 - (\beta + \gamma))G$. Then the weighted aspect vector is trimmed to retain only the top n = 30 aspects, number which gives the best precision-recall trade-off in the Wikipedia-based evaluation.

6.2 Wikipedia vs. Other Web Biographies

To establish an expectation for the range of values, we first compared Web biographies against Wikipedia. For 10 random development topics, we manually picked 'the best' biographical page from the top 10 Web search results. We then compared the sentences in those biographies against the sentences in Wikipedia. Figure 4(a) shows the results for this comparison. All sentence-level metrics indicate a lexical mismatch between the Web biographies and Wikipedia. Concept-based precision and recall measures are also low.

6.3 Comparison of Selection Methods

We first determined the parameter values for each sentence selection method by using exhaustive grid search on the development set and then we compared the results.

The performance numbers for the automatic sentence selection methods (Figure 4(b)) are comparable to those obtained for the Web biographies. As expected, Full Aspect obtains poor performance on the diversity based measures. Novelty and Diversity, which use Full Aspect for initial ranking, do not provide substantial gains, due to the poor-quality initial ranking. For Diversity, removing aspects from the aspect vector leads to poor quality sentences being retrieved as the context for ranking gets reduced. Typical and D-T outperform the other sentence selection methods across all measures. Typical focuses on retrieving the best possible sentence for each aspect by leveraging the aspect specific context. Conversely, D-T improves the concept-level precision and recall measures, as explicitly promoting diversity improves D-recall. Even though the aspect vector is trimmed at each iteration, D-T is able to handle the reduced context better than Diversity because it interpolates the aspect vector and the aspect-specific context vector.

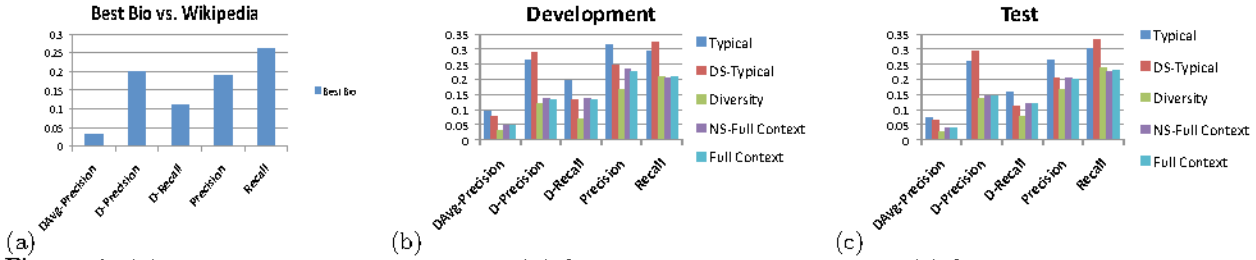
Based on these findings, we chose D-T with empirically-best $\delta = 0.5$ and $\lambda = 0.25$, for sentence selection in our final system. Also, we obtained the corresponding aspect models interpolation parameters as being: $A_m = 0.1S + 0.7R + 0.2G$.

7. SENTENCE ORDERING

Once sentences pertaining to the important aspects of a given topic are collected, it is desirable that they are presented in a coherent manner. Biographies typically follow the natural timeline of events in a person's life. Because Wikipedia biographies usually obey this rule, we use them as training data to learn how to order biographical sentences. Our approach is to assign precedence scores for pairs of words based on Wikipedia evidence and then use these scores to order the sentences collected for each topic.

We first build a restricted vocabulary of words with at least 5 occurrences in the Wikipedia training set. For each

[4] We exclude Wikipedia articles from the pool of Web documents retrieved.



(a) (b) (c)

Figure 4: (a) Web bio vs. Wikipedia pages; (b) Sentence selection on dev set; (c) Sentence selection on test set

pair in this vocabulary, we count the number of times one word was used in a sentence preceding a sentence containing the other word. Then, we assign a precedence score for each pair of sentences by combining these word-based statistics. To produce the final ordering of sentences, we compute for each sentence an overall precedence score against all other sentences by aggregating the pairwise scores, and sort the sentences in the decreasing order of overall scores.

We experimented with both binary and frequency-based combination methods, with and without back-off, for sentence pairwise precedence, as well as two schemes (averaging and product of likelihood ratios) for overall scoring.

We evaluated our scoring methods on the 100 development topics, by extracting individual sentences from each corresponding Wikipedia articles and then attempting to recover the original ordering of those sentences. Despite their simplicity, our methods achieved high Spearman rank correlations for this task (0.55-0.65). In particular, we found the frequency-based pairwise scoring in conjunction with product of likelihood ratios to perform the best.

8. TOPIC PAGE EVALUATION

The automatic evaluation of our final system on the test set shows results consistent with those obtained during development (Figures 4(c) and 4(b), respectively).

Additionally, we manually evaluated the aspect models and topic pages generated by the final system for a set of 20 random topics from our test collection (Table 3). We limited the number of sentences in the topic pages to 20 in order to match the size of typical Web search result pages. The evaluation was done independently by two annotators, who first read the Wikipedia page for each topic and extracted a set of aspects covering personal life, and career facts. Subsequently, the aspect models and the generated topic pages were evaluated based on the knowledge gathered from Wikipedia. When the information retrieved appeared to be new, other Web sources were consulted. We opted for a 3-point scale {0, 0.5, 1} to limit subjectiveness but allow some degree of uncertainty and granularity of relevance.

For aspect models, we obtained an average precision of 0.33 for self aspects and 0.30 for related aspects. The annotator inter-agreement rate was 86%, with a Kappa of 0.66.

The topic summaries were evaluated on multiple dimensions, as follows: *precision* (is the information important to the topic and correct?), *grammaticallity* (is the information conveyed accurately?), *non-redundancy* (how much new information is conveyed on average by each sentence?), *novel information versus Wikipedia* (are facts not covered in Wikipedia presented?), and *recall* (how well the aspects from Wikipedia are covered in the summary?).

Table 4 shows the results obtained for the 20 test topics by macro-averaging the two annotators' scores. The grammaticallity, non-redundancy, and novelty scores were aggregated only for sentences with a non-zero precision score. Most topics (14) scored over 0.5 in precision. We observed very good inter-annotator agreement, of 89% at sentence level, with a correspondingly high Kappa coefficient of 0.77. Precision is much higher for topic pages than for the aspect models, which indicates that the retrieval/selection stage of our system is able to tolerate noise in the aspect models.

To verify that our system succeeds in differentiating itself from the current search approaches, we also performed a comparative evaluation with Bing. We first compared the generated topic pages and the search result pages globally, in terms of relevant information made available to the user. For 15 out of 20 topics, we strongly preferred the topic pages, in 4 cases, the information provided was comparable, while in one case, the search engine result page was more informative. We then judged the search engine results by using the same guidelines devised for topic pages. Our system substantially outperformed the Bing baseline on all metrics (Table 4).

Table 3: List of topics used in the manual evaluation.

Bette Midler	Harvey Keitel	Mario Cuomo	Newt Gingrich
Billy Bragg	Holly Hunter	Mario Lemieux	Reese Witherspoon
Bob Brady	Joe Theismann	Marion Jones	Roberto Benigni
Carmen Electra	Julie Walters	Matthew Santos	Saxby Chambliss
Elton Brand	Lindsey Graham	Monica Lewinsky	Sean Young

Table 4: Macro-averaged performance for 20 test topics.

	Prec.	Gramm.	Non-Red.	Novelty	Recall
topic pages	0.50	0.83	0.93	0.29	0.53
Bing	0.37	0.61	0.57	0.07	0.30

9. CONCLUSION

We investigated the automatic generation of topic pages for biographical topics and we presented a general framework for this task. Our evaluation indicates the viability of automatic generation of topic pages as an alternative to the current search-based exploration of the Web.

10. REFERENCES

[1] H. Alani et al. Automatic ontology-based knowledge extraction from Web pages. *IEEE-IS*, pages 14–21, 2003.

[2] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. *SIGIR*, pages 314–321, 2003.

[3] F. Biadsy, J. Hirschberg, and E. Filatova. An Unsupervised Approach to Biography Production using Wikipedia. *ACL-HLT*, pages 807–815, 2008.

[4] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. *EMNLP-CoNLL*, pages 708–716, 2007.

[5] E. Filatova and J. Prager. Learning occupation-related activities for biographies. *HLT-EMNLP*, pages 49–56, 2005.

[6] F. Lacatusu et al. LCC's Gistexter at DUC'06: Multi-strategy multi-document summarization. *DUC*, 2006.

[7] A. Nenkova and A. Louis. Identifying correlates of input difficulty for generic multi-document summarization. *ACL-HLT*, pages 825–833, 2008.

[8] A. Nenkova, L. Vanderwende, and K. McKeown. A compositional context sensitive multi-document summarizer. *SIGIR*, pages 573–580, 2006.

[9] M. Pasca. Turning Web Text and Search Queries into Factual Knowledge. *AAAI*, pages 1225–1230, 2008.

[10] D.R. Radev and K. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics 24(3)*, pages 469–500, 1998.

[11] B. Schiffman, I. Mani, and K.J. Concepcion. Producing biographical summaries. *EACL*, pages 450–457, 2001.

[12] H.J. Zeng, Q.C. He, Z. Chen, W.Y. Ma, and J. Ma. Learning to cluster web search results. *SIGIR*, pages 210–217, 2004.