Topic Pages: An Alternative to the Ten Blue Links

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

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

***Abstract*—We investigate the automatic generation of *topic pages* as an alternative to the current Web search paradigm. Topic pages explicitly aggregate information across documents, filter redundancy, and promote diversity of topical aspects. We propose a novel framework for building rich topical aspect models and selecting diverse information from the Web. In particular, we use Web search logs to build aspect models with various degrees of specificity, and then employ these aspect models as input to a sentence selection method that identifies relevant and non-redundant sentences from the Web. Automatic and manual evaluations on biographical topics show that topic pages built by our system compare favorably to regular Web search results and to MDS-style summaries of the Web results on all metrics employed.**

***Keywords*-Web search; topic page; query log; aspect model.**

I. Introduction

Web search results are usually presented as ten blue links along with short snippets that serve as individual summaries of the retrieved documents. As noted by Clarke et al. [1], the quality of these *search snippets* affects the users' perception of relevance and influences their click behavior. However, because search snippets are short and generated independently of each other, they provide only superficial and possibly redundant coverage of topical aspects.

As an alternative to the ten blue links paradigm, we propose the automatic generation of *topic pages*. We envision topic pages to aggregate and organize information on different aspects of a topic, with pointers to information sources. Figure I(a) shows an example topic page automatically generated by our system for the topic "William Shatner". It covers different aspects relevant to William Shatner, such as acting career, famous movies, books, and even recent TV commercials. The page also provides links to various Web sources for additional information on each aspect of the topic. In contrast, as shown in Figure I(b), the search snippets cover only a small set of these diverse aspects. By explicitly addressing redundancy and diversity, topic pages can provide an useful alternative to the search-based exploration of the Web for a large set of Web queries.

Automatically generating such topic pages raises several interesting challenges. We focus on two of them: 1) identifying and assembling diverse aspects pertaining to a topic, and 2) retrieving and organizing information corresponding to these diverse aspects in a non-redundant fashion.

Multi-document summarization (MDS) systems use notions of centrality and novelty to tackle similar challenges [2], [3]. However, direct application of MDS models to build topic pages from the Web search results for a topic is challenging because of the vast number of retrieved Web pages, the diverse document layouts, as well as the prevalence of out-of-topic information in many of those pages. While limiting the set for summarization to only several top ranked Web documents that follow a certain layout helps in reducing noise to some extent, it also reduces diversity of information and limits the number of sources used in a topic page.

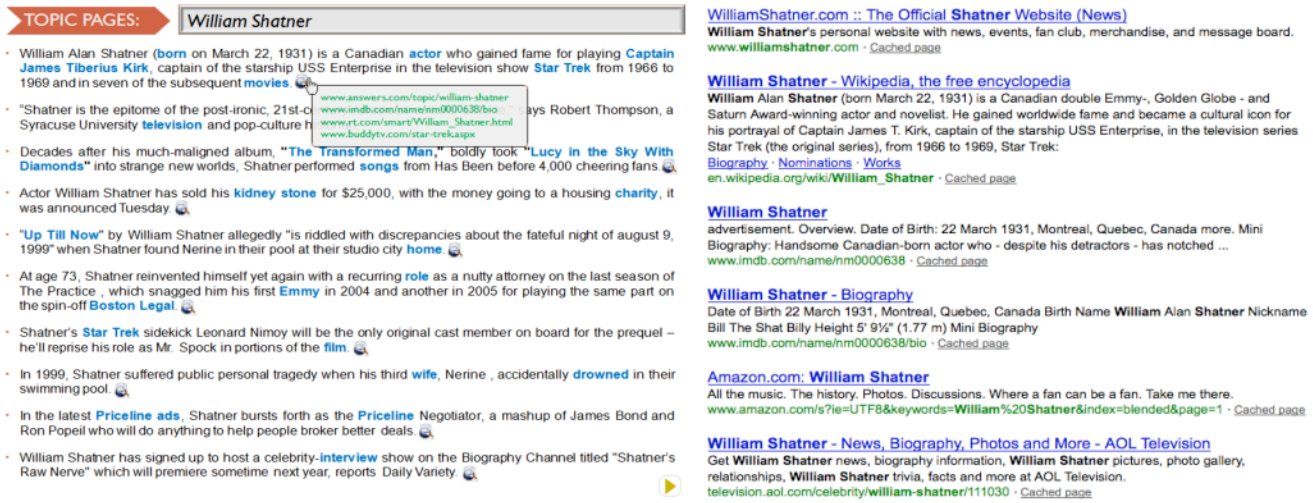
We investigate a direct two-tiered approach to generation of automatic topic pages, which does not rely on implicit discovery of topics in the summarized documents. Instead, we first identify diverse topical aspects and then organize information pertaining to these aspects in order to maximize relevance and diversity. We leverage Web query logs, which aggregate the information needs of Web search users with respect to many topics. We build query-log-based *aspect models* that capture a consensus of user interests with respect to the target topics. We then gather and organize information from the Web through sentence selection techniques that explicitly enforce relevance and diversity.

We conduct automatic and manual evaluations of our general approach on a set of topics related to people, such as actors and politicians, using Wikipedia as a reference collection. Our evaluations attempt to provide evidence for the viability of automatic topic pages as an alternative to the traditional ten blue links paradigm for Web search.

II. Related Work

A. Multi-document Summarization

Topic page generation can be viewed as a topic-focused multi-document summarization task [4]. However, the primary goal of MDS is to summarize a given set of input documents whereas, the main goal for topic page generation is to retrieve diverse information pertaining to topical aspects. This difference raises two issues that limit the applicability of typical MDS techniques for topic pages.



(a) Automatically generated topic page for "William Shatner"

(b) Web search result page for "William Shatner"

Figure 1. Examples of topic page and result page from a major Web search engine. By focusing on non-redundant presentation of various aspects of interest to users as captured in Web search logs, topic pages can expose rich information about the topic directly while still serving as Web content hubs.

First, typical MDS systems only utilize lexical properties of the input documents to determine sentences that need to be added to the summary, as noted by Nenkova et al. [5]. For example, Radev et al. [3] use a centroid based approach, and Erkan and Radev [2] use centrality of sentence nodes in a lexical graph of the documents to select sentences. Even more sophisticated approaches rely on the input documents to determine the themes of the topic [6], or build document content based topic models [7]. Instead of relying on a seed set of documents to derive richer models for document representation, we explicitly determine important aspects for a topic and extract sentences from the Web that cover these aspects. Our approach is similar to that of [8], which proposes the use of Wikipedia to directly rank sentences. However, we gather topical aspects by aggregating queries issued by Web search users and design sentence selection methods that explicitly enforce coverage of these aspects.

Second, non-cohesive input documents are hard to summarize [9]. Even though documents in Web search results are often on-topic, they frequently contain other irrelevant information, lack cohesiveness, or provide noisy coverage of one particular aspect. To overcome such problems, Lacatusu et al. [10] employ a multi-strategy system that breaks down a complex query into simpler queries and produces summaries as answers to these queries. In a similar fashion, we generate aspect models for a topic by employing query logs and use them to construct the topic pages, thereby avoiding the difficult task of summarizing non-cohesive Web documents.

B. Biography Generation

Early work on biography generation has focussed on multi-document summarization of information present in news collections [11], [12]. Alani et al. [13] use pre-defined biographical templates, with aspects such as "painting style", and use information extraction techniques to collect biographical facts in order to fill-in the artist templates. Filatova and Prager [14] predict the occupation of a person by identifying person-specific, occupation-specific, and general events for biographies. [15] analyze several approaches for extracting characteristic facts from biographies, including contextual patterns and positional distribution of certain types of facts. Biadsy et al. [16] learn a biographical sentence classifier from Wikipedia and the TDT4 corpus. Sauper and Barzilay [17] also employ Wikipedia to derive domain-specific templates by using the most frequent section headings. These templates are then populated with text segments extracted from Web documents retrieved for queries composed of the topic and the section headings. A key difference of our approach is that it uses rich interpolated aspect models with various degrees of specificity derived from Web-search logs as opposed to employing a predetermined set of aspects [13], utilizing a Wikipedia-based biographical sentence model [16], or deriving occupational templates by using section headings from a preexisting corpus of in-domain documents [17]. The model richness allows us to query the Web for topic-specific aspects and retrieve sentences from multiple sources for each individual aspect. Additionally, by employing aspects queried for the target topic as well as commonly queried for similar topics by Web search users, the proposed system can capture automatically multiple occupational roles of the same person.

C. Aspect Models

Pasca [18] proposes a seed-based framework for weakly-supervised class attribute extraction and attribute propagation in conceptual hierarchies, populated from query logs and Web documents. While such a framework could be employed in building aspect models for topic pages also, we opted for aspect models with lower complexity and easier to generate/evaluate.

D. Organizing Web Search Results

Daume and Brill [19], Zeng et al. [20], and Cheng et al. [21] investigate Web search result clustering, which implicitly attempts to discover sub-topics within the search results. Zhuang and Cucerzan [22] and Wang and Zhai [23] explicitly identify aspects relevant to a topic from query logs, but use them to re-rank the top search results and to guide the clustering of search results for the topic query, respectively. In contrast, we use the aspects derived from query logs to directly retrieve Web sentences and organize them in a topic page with pointers to Web sources.

Major Web search engines have tackled the idea of building topic pages by allowing contributors to provide extensive documents on a topic (e.g., Google Knol [24]), by generating them based on predefined templates (Yahoo! Glue [25]), or by performing a "deep analysis of what the Web has to say" about various aspects of a topic (Lycos Labs' Retriever system [26]).

III. Overview

A. System Architecture

Our framework for automatically generating topic pages comprises three main components: (a) aspect extraction, (b) content retrieval/selection, and (c) content organization. First, we extract a diverse set of aspects relevant to the topic from search query logs. Then, we attempt to find sentences that cover these aspects meaningfully. Finally, we organize these sentences into a non-redundant readable summary of the topic. In this paper, we focus on aspect extraction and content retrieval/selection.

B. Experimental Setup and Data Collection

We use Wikipedia to create a reference collection of biographical topics to develop, train, and evaluate the automatic topic page generation system and its subcomponents. Wikipedia is generally a good reference because it often provides exhaustive coverage of topics. Additionally, we use other Web resources when necessary to account for new information generated by our system.

To select biographical topics, we first gathered Wikipedia pages that are labeled with the category "Living people". We use the Wikipedia page titles as topic names, after removing parentheticals. Less than 5% of these names are duplicates that indicate ambiguity problems. Our preliminary experiments show that such ambiguity can be handled well by employing disambiguation systems trained on Wikipedia [27], [28] to hypothesize the appropriate topic for each Web page retrieved. Nonetheless, to simplify the experimental setup, we eliminated these topics from our collection.

In order to create a diverse pool of topics, we collected category labels from a small set of known topics such as sports politics, acting, and music. Then, we gathered all pages assigned to these category labels. To avoid topics with very little user interest, we also removed topics with fewer than 20 entries in a six-month query log of 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 (available at: http://research.microsoft.com/en-us/people/silviu/topics/data).

Table I. Percentage of instances for which users queried the exact topic name vs. the topic and at least one aspect during two consecutive months (queries submitted less than 10 times during each month were discarded).

Topic	Exact	+ Aspect	Exact	+ Aspect
William Shatner	61%	39%	59%	41%
Oprah Winfrey	24%	76%	50%	50%
Bill Clinton	48%	52%	72%	28%
George Clooney	26%	74%	61%	39%
Daniel Radcliffe	51%	49%	42%	58%
Lebron James	66%	34%	60%	40%

IV. Aspect Models

Our goal is to build topic pages that provide concise information on the most important aspects of a topic with respect to Web users' needs (with pointers to Web sources that address those aspects). Most multi-document summarization techniques avoid explicit identification of aspects by relying on document structure and term statistics in the set of documents to be summarized. However, in our setting it is unclear how to gather a collection of Web documents that would result in a desirable topic page by applying such techniques. Instead, we attempt to identify explicitly the most important aspects for a topic by mining Web search query logs, which can be viewed as aggregating the interests of a vast number of Web users.

As illustrated in Tables I and II, users often submit queries that contain other words in addition to the topic name in order to retrieve information relevant to one particular aspect of the topic. The observed distribution of queried aspects can be heavily biased towards events occurring within the time frame of the analyzed logs. This distribution is biased further by the lexical diversity in queries that refer to the same aspect (e.g., "dead", "drowning", and "death" are used in conjunction with "wife" for the topic "William Shatner"). Our experiments showed that while term clustering for each topic addresses aspect redundancy for some cases, it also frequently groups together terms for distinct aspects while assigning terms corresponding to one aspect to multiple clusters. While it is essential for topic pages to cover aspects that correspond to the users' interests, even if of temporary nature, it is also desirable to include generally-important information about the topic. This premise is supported by the substantial percentage of user queries that only contain the topic name, which may indicate demand for general information or willingness to read about some specific aspect in a broad-coverage topic context.

Table II. THE MOST FREQUENT 10 USER QUERIES THAT CONTAIN "WILLIAM SHATNER" AS AGGREGATED OVER SIX MONTHS OF LOG DATA AND OVER TWO EXTRA CONSECUTIVE MONTHS.

6 previous months	1st month	2nd month
... character	... game show	... game show
... biography	... wife murdered	... gun control
... wife	... gun control	comedy central ... roast
... movies	... novel	... biography
... bio	... commercial	... wife
... commercials	... biography	... wife murdered
... news	... spokesman	program starring ...
roast of ...	... travel	... sings 1978 sci fi awards show
... music	... mask	... character
... sings	... wife	... travel

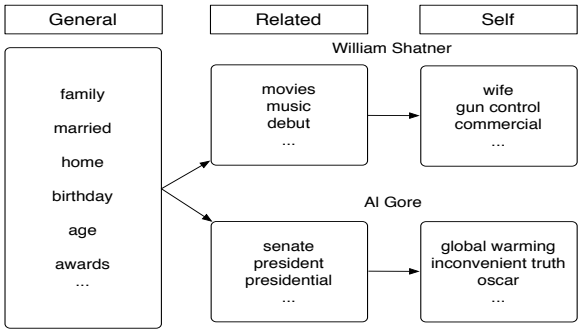


Figure 2. Aspect models for "William Shatner" and "Al Gore".

A. Types of Aspects

To lessen query-log biases and to capture a diverse set of aspects important for the type of topic targeted, we examine an approach that combines three types of aspects: *self*: aspects specific to the topic, *related*: aspects common across related topics, and *general*: aspects relevant to all topics. Figure 2 exemplifies the three types of aspect models, as built from query logs for "William Shatner" and "Al Gore".

Self Aspect Models: We construct the self aspect model for a topic by first extracting queries that contain the topic of interest. Then, we select the most frequent n terms that occur in those queries after filtering out stopwords such as prepositions and determiners. To capture multi-word concepts, we tokenized the queries by using a list of names and concepts mined from Wikipedia. Thus, the terms in our aspect models can be either individual words or phrases.

Related Aspect Models: 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.

General Aspect Models: Finally, we build a general aspect model by combining the self aspect models of all the Wikipedia topics in our pool.[1] Additionally, since this model is generated only once, we filter high-frequency but biographically-meaningless terms such as "pics" and "official page" from the model, as such terms appear spuriously across all topics and can rank higher than the informative terms in the aggregation. We set the size of this model to the same value n.

B. Experiments. Comparison to Wikipedia

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. For the comparison, we use the unique Wikipedia page titles as concepts and the anchor texts of the Wikipedia links as the vocabulary for these concepts. Based on the comparison, we find that the recall of the Wikipedia concepts is very low around 4% for self, and 6% for related, mostly because the Wikipedia pages often provide exhaustive biographical coverage, while query logs only capture those aspects of interest to search engine users. 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). Also, we find that the related aspects achieve better overall precision and recall compared to the self-aspect models. We believe it is due to the temporal bias in the self aspect models. While this automatic comparison is inherently biased against new concepts and does not account for lexical mismatches, it provides a rough quality estimate for the extracted aspects, which we use to tune the parameters of the models.

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

V. SENTENCE SELECTION

We now address the problem of selecting sentences from the Web that cover a given set of topical aspects. The main challenges in selecting sentences from the Web lie in handling ungrammatical sentences, identifying relevant sentences for each aspect, and covering the diverse aspects pertaining to the topic.

A. Grammaticallity

Sentences extracted from Web documents can be ungrammatical because of html parser failures, sentence boundary detection failure, or unreliable content such as blog postings and user responses.

To identify ungrammatical sentences from the candidate sentences pool, we use a logistic regression classifier that employs html features (such as hyperlinks and html tags), lexical (such as unigram likelihood and perplexity of sentences), and orthographic features (such as the number of special characters and upper to lower case ratio). We manually labeled a set of 100 grammatical and ungrammatical sentences extracted from the Web for the training topics. On another small test set, this classifier achieved more than 80% precision at 85% recall in identifying grammatical sentences.

Table III. EXAMPLE SENTENCES FOR THE TOPIC "WILLIAM SHATNER" AND THE ASPECT "WIFE".

During his speech, **Shatner** mentioned that he and his **wife** are chiropractic advocates and have been scanned with CLA's [...]
Shatner told fire officials that the last time he saw his **wife** alive was early Monday morning.
Shatner's third **wife** Nerine Kidd drowned in the swimming pool of their Los Angeles home on August 9, 1999.
In 1999, **Shatner** suffered public personal tragedy when his third **wife**, Nerine, accidentally drowned in their swimming pool.

B. Relevance

To select sentences that cover the aspect (or aspects) of interest, we can retrieve sentences that contain both the topic and the aspect. However, several sentences on the web can contain both the topic and the aspect of interest and not all sentences capture the relevant connection between the topic and the aspect. For example, Table III shows four sentences retrieved from the Web for the topic "William Shatner" and the aspect "wife". Even though the first two sentences contain both the topic and the aspect, they do not capture the typical connection between them.

Further, we find that the lexical contexts in which an aspect occurs usually differ from both the contexts of other aspects and the general topic context, as seen in the last two sentences of Table III. Therefore, to select sentences that best capture the aspects of interest, we build aspect specific contexts and the overall topic context.

To build the aspect-specific context vector for a given aspect, we aggregate the terms and their frequencies in sentences that contain both the topic and the aspect. For the overall topic context, we normalize all the aspect-specific context vectors and aggregate them into a single topic vector.

Finally, for each aspect, we rank candidate sentences using their vectorial similarity to the aspect-specific context vector and the overall topic context, as described further in Section V-D. For example, the context vector for "William Shatner" and the aspect "wife" has high weights for "swimming pool", "Nerine", "drowned", "1999", and "tragedy", which lead to the selection of the fourth sentence in Table III.

C. Diversity

Aspect synonymy and selection of sentences that contain multiple aspects can lead to redundancy and reduced diversity of aspects covered in the summary. Typical novelty techniques do not work well for promoting diversity, as they do not ensure the coverage of all the aspects of interest. To promote diversity, we employ the following process, which can be applied to any of the sentence selection techniques: We first select a sentence that has the highest similarity with a given aspect vector, then we remove the aspects that the sentence covers from the aspect vector. We repeat this process until no aspects are left.

D. Experiments

We employed the development set to conduct sentence selection experiments and tune the parameters of our system. During training, each method is allowed to learn its own weights for interpolating the self (*S*), related (*R*) and general (*G*) aspect models: $A_m = \beta S + \gamma R + (1 - (\beta + \gamma))G$. The weighted aspect vector is then trimmed to retain only the top n = 30 aspects, number which gives the best precision-recall trade-off in the Wikipedia-based evaluation.

1) Methods: We compare five sentence selection approaches including our approach for addressing relevance and diversity. To perform sentence selection, we create a candidate pool of sentences. For each topic, we create queries by combining the topic and the aspect and issue them to the Bing search engine to retrieve web pages. These web pages are then processed to extract a pool of candidate sentences. In all cases, we first remove ungrammatical sentences from the candidate pool using the sentence classifier, then we train the parameters for the selection method using exhaustive grid search. We detail the five methods for selecting a set of sentences I. For all methods, we begin with the candidate sentence pool, S, and the aspect model A_m. We limit the number of selected sentences ($|I| = n$) to 30.

The five methods investigated are:

i. Full-context: Rank sentences based on their cosine similarity to the full aspect vector, $f(s) = cosine(s, A_m)$, and add the top n sentences to I.

ii. NS-Full-context: Rank sentences by using the Full-context method and select the top-ranked sentence. Iteratively, select new sentences based on a linear combination of their original score (full-context score) and their dissimilarity with the currently selected sentence set (I).

I $= \arg\max_{s \in S} f(s)$
until $|I| < n$ do
(1) $n(s) = \lambda * f(s) + (1 - \lambda) * \min_{s_i \in I}(1 - cosine(s, s_i))$
(2) $s^* = \arg\max_{s \in S} n(s)$
(3) $I = I \bigcup \{s^*\}; S = S - \{s^*\}$

iii. Diversity: Start with the full aspect vector A_m. Iteratively, select the top ranked sentence based on its cosine similarity to A_m. Then, down-weight the aspects in A_m that are covered by the selected sentence. Repeat this process until the desired number of sentences are selected.

until $|I| < n$ do
(1) $s^* = \arg\max_{s \in S} f(s)$
(2) $I = I \bigcup \{s^*\}; S = S - \{s^*\}$
(3) $A_m = A_m - \delta * A_s$

iv. Typical: For each aspect, extract sentences that contain the aspect and the topic (T_a), and build the term context vector (V_a). Rank all sentences based on their cosine similarity with C_a, a linear interpolation of the two vectors.

for each $a \in A_M$ do
(1) $V_a = \sum_{s \in T_a} s$
(2) $C_a = \lambda * A_m + (1 - \lambda) * V_a$
(3) $\forall s \in S, t(s) = cosine(s, C_a)$
(4) $I = I \bigcup \arg\max_{s \in S} t(s)$

v. DS-Typical: Start with the full aspect vector. Select an aspect from this vector. Use the Typical method to get the best candidate sentence for the aspect. Then, remove all the aspects that were covered by the selected sentence. Repeat the process until no more aspects remain in the vector or if the desired number of sentences are selected.

2) Metrics: To evaluate the different methods, we use sentences in Wikipedia pages as reference sentences for each topic. We employ both term-based metrics (precision and recall), which measure the term overlap in the selected sentences as a whole, and sentence-level metrics, *D-Precision, D-Recall* and *D-Average precision*, which measure sentence overlap between the selected set and the corresponding Wikipedia page for each topic. The *D*-metrics are basically modified versions of their standard counterparts that favor diversity in the set of selected sentences. To compute these *D*-metrics, we assume that two sentences match each other if their term-based cosine similarity is greater than 0.7; also, once a Wikipedia sentence is matched we remove it from further consideration.

3) Results: The absolute values obtained in these development experiments (Figure 3(a)) were low for all metrics. Therefore, to establish a range of values expected for this type of evaluation, we compared Web biographies against Wikipedia pages in the same manner. For 10 randomly chosen development topics, we manually picked the best biographical page from the top 10 Web search results returned for the topics used as queries. The results obtained for these experiments were similarly low, as shown in Figure 3(c).

The contrastive analysis of the five sentence selection methods investigated shows that the novel methods we propose, Typical and DS-Typical, outperform the baseline Full-Context and NS-Full-Context methods in both development and final evaluation experiments (Figures 3(a) and 3(b)).

As expected, the Full-Context method, which does not address novelty and diversity, obtains poor performance on the diversity based measures. Even the NS-Full-Context and Diversity methods, which use Full-Aspect for initial ranking, do not provide major gains. This is mainly due to the fact both methods are affected by the poor quality of the initial ranking. Additionally, for Diversity, removing aspects from the aspect vector leads to poor quality sentences being retrieved as the context for ranking is continually reduced.

Typical focuses on retrieving the best possible sentence for each aspect by leveraging the aspect-specific contexts. DS-Typical improves the concept-level precision and recall measures as expected, by explicitly promoting diversity. Even though the aspect vector is trimmed in each iteration, DS-Typical is able to handle the gradually reduced contexts better than the simple Diversity method by interpolating the overall contexts and the aspect-specific contexts to score sentences in each iteration. This can be seen as a trade-off between *D*-precision versus *D*-recall.

Table IV. 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

Based on these findings, we chose to employ DS-Typical with the empirically-best $\delta = 0.5$ and $\lambda = 0.25$ for sentence selection. Among the methods compared, it generates the best set of diverse, non-redundant sentences that cover aspects relevant to the topic. Furthermore, as a result of these experiments, we also obtained the corresponding set of combination parameters for the aspect models: $A_m = 0.1S + 0.7R + 0.2G$, which indicates that the related models built from aspects common to a small set of related topics capture most of the information important for topic summaries.

VI. TOPIC PAGE EVALUATION

In addition to the automatic Wikipedia-based evaluation, on which our approach shows consistent improvements over baselines, we conduct a manual evaluation on a set of 20 randomly chosen test topics (Table IV). We compare automatic topic pages, generated with parameters learned during training, to the results page of the Bing search engine and to a multi-document summarization system. To keep the summary sizes comparable to those of typical search results pages, we limit the number of sentences in our summaries to 20.

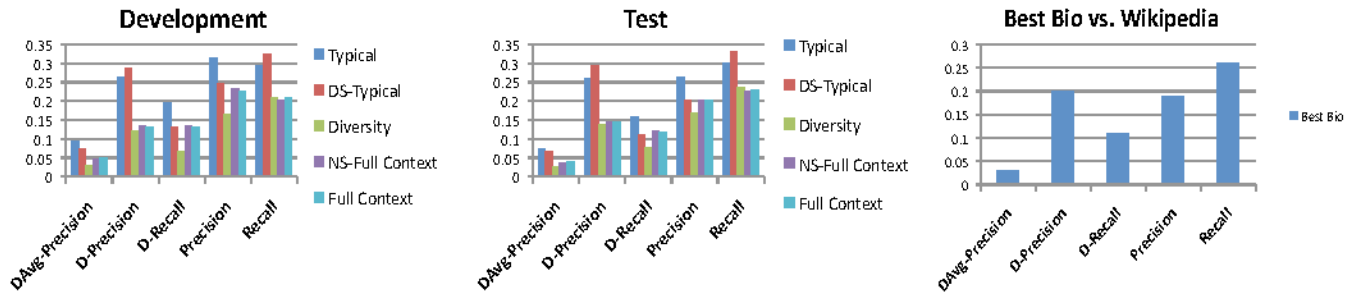
A. MDS Baseline

We use LexRank [2], a state-of-the-art multi-document summarization system, which is available as part of the MEAD software[2], version 3.11. We apply the identical sentence segmentation and remove ungrammatical sentences before summarization. LexRank is a rather computationally intensive algorithm (polynomial in the number of sentences in the documents). Even when we limit the search results to the top 30 documents, LexRank takes about 30 minutes to complete the summarization.

B. Guidelines

The evaluations were done independently by two annotators as follows: For each topic, the annotators read the corresponding Wikipedia page and extracted a set of (subjectively) important aspects covering personal life, career, and trivia facts. Depending on the perceived richness of the topic, the annotators identified between 10 and 20 such aspects. Then, the annotators evaluated each sentence in the summary on five criteria: 1) precision: the importance and accuracy of the aspect information with respect to the topic; 2) grammaticallity: whether the information was conveyed effectively, without causing difficulties in resolving references;

[2]http://www.summarization.com/mead



(a) Development set (b) Test set (c) Web biography

Figure 3. Automatic evaluation results by using Wikipedia as gold standard.

3) non-redundancy: whether the information in a sentence is new with respect to the preceding sentences; 4) novelty: whether the information is new and not covered by the current Wikipedia page on the topic; 5) recall: the number of manually extracted Wikipedia aspects covered by the whole summary. We used a 3-point scale {0, 0.5, 1} instead of binary {0, 1} to allow some degree of uncertainty and granularity of relevance, as we noticed in our development experiments that numerous sentences fall in between being perfect and detrimental, as they contain partial, extra, or slightly redundant information, for example.

C. Results

Table V shows the scores obtained on the 20 test topics by macro-averaging the two evaluators' scores (i.e., we first compute scores for each topic and then aggregate those to obtain average and standard deviation numbers for the whole set). The grammaticallity, non-redundancy, and novelty scores were aggregated only for those sentences which received a non-zero precision score.

The proposed DS-Typical model substantially outperforms both the Web search result pages and the MDS baseline on all metrics. Clearly, the Web search engines are not optimized for this type of evaluation and the MDS systems are not well-suited for the task of summarizing Web search results. As an indicator of the absolute quality of the topic pages, we note that DS-Typical scores over 0.5 in precision for most topics (14), and only 2 topics had very low precision values (between 0.1 and 0.25). The macro-averaged recall of 0.53 is much higher than the recall numbers observed in the automatic evaluations. The annotator inter-agreement rate for this task was high, around 89% at sentence level, with a correspondingly high Kappa coefficient of 0.77.

Additionally, we compared the topic pages, the LexRank summaries, and the search result pages globally, in terms of overall relevant information made available to the user with respect to each topic. For 15 out of 20 topics, the topic pages were preferred to the Web search results, in 4 cases the information provided was comparable, while in one case, the search result page was judged as more informative. Against LexRank, topic pages were preferred for 17 out 20 topics, and judged as being similar in content in 3 cases.

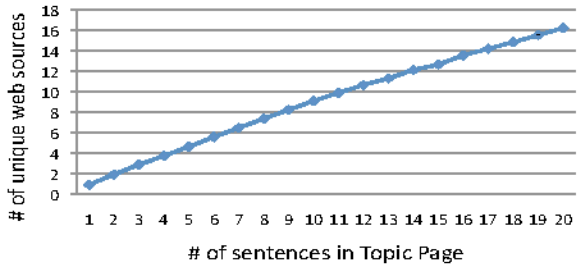


Figure 4. Unique Web sources in the generated topic pages.

Using the same evaluation criteria, we also evaluated the generated aspect models. We obtained an average precision of 0.33 for the self aspects and 0.30 for the related aspects.[3] The annotations show an inter-agreement rate of 86%, with a Kappa coefficient of 0.66.

Note that the average precision score for the topic pages (0.50) is much higher than the precision scores obtained for the aspect models (0.30 and 0.33), which indicates that the overall system is able to tolerate noise in the aspect models through sentence selection.

Finally, we analyzed the average number of Web sources used for summaries of various lengths. As shown in Figure 4, the generated summaries contain topic sentences from diverse Web sources (more than 16 sources for summaries of 20 sentences), and thus, are suitable for use as an alternative way for exploring the Web for biographical information to the current Web search engines.

VII. Discussion

To understand differences with a state-of-the art biography generation system, we obtained Wikipedia-like summaries automatically generated by the system of [17] for the six

[3]Since the number of aspects in all models is the same, micro-averaging and macro-averaging results are identical.

Table V. MACRO-AVERAGED PERFORMANCE FOR 20 TEST TOPICS.

Method	Precision	Gramm.	Non-Redund.	Novelty	Recall
Web Search Snippets	0.37 (0.12)	0.61 (0.22)	0.57 (0.23)	0.07 (0.09)	0.30 (0.16)
LexRank	0.22 (0.11)	0.61 (0.18)	0.71 (0.26)	0.19 (0.20)	0.34 (0.21)
Proposed System (DS-Typical)	0.50 (0.17)	0.83 (0.13)	0.93 (0.15)	0.29 (0.17)	0.53 (0.19)

Table VI. TOPICS RELATED TO A TARGET TOPIC AS DERIVED AUTOMATICALLY FROM AGGREGATED WEB SEARCH SESSION DATA.

Target topic	Related topics
bank of america	capital one, mbna, citibank, first usa
fox news	cnn, msnbc, bbc, usa today
priceline	expedia, orbitz, hotwire, travelocity
delta	united, northwest, continental, america west

actor topics[4] in our test set.[5] While this set is too small for a meaningful quantitative comparison, it gives insights into the paradigmatic differences. The Wikipedia-like summaries provide in-depth information extracted from a small set of sources (around 4 per topic), while our topic pages use a larger number of sources (on average, 15 per topic) to cover more diverse information but in less detail. For example, for "William Shatner", the Wikipedia-like summary covers five different movies and his stage career, whereas the topic page mentions only his most famous movie but covers aspects of interest to Web search users related to his other occupational roles (commercials, books, songs), which are not mentioned in the Wikipedia-like summary.

While our study focused on biographical topics, the techniques employed are applicable to other types of topics, as our approach for generating aspect models is not domain-specific. Some preliminary results for deriving *related* topics by using search session information are shown in Table VI. More sophisticated techniques for class-attribute extraction from query logs could further improve both the quality and generalizability of aspect models to other topic types. The investigated sentence selection techniques are also applicable to other types of topics, as they are based on general notions of relevance, diversity and novelty.

VIII. CONCLUSION

We investigated the viability of automatically generated topic pages as an alternative to the search-based exploration of Web content for biographical topics. We presented a general framework for this task, which includes methods for identifying aspects that capture user needs with respect to a topic of interest and methods of sentence selection that explicitly account for relevance, diversity, and reduced redundancy of information between sentences. The implemented system outperformed substantially both baselines employed (Web search pages and LexRank).

[4]This system is currently trained for actors only.

[5]We would like to warmly thank Christina Sauper for kindly running experiments and providing the output to us.

REFERENCES

[1] C. Clarke, E. Agichtein, S. Dumais, and R. White, "The influence of caption features on clickthrough patterns in web search," in *Proceedings of SIGIR '07*, 2007, pp. 135–142.

[2] G. Erkan and D. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, no. 2004, pp. 457–479, 2004.

[3] D. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing and Management*, vol. 40, no. 6, pp. 919–938, 2004.

[4] H. Dang, "Overview of DUC 2005," in *Proceedings of DUC 2005*, 2005.

[5] A. Nenkova, L. Vanderwende, and K. McKeown, "A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization," in *Proc. of SIGIR 2006*, 2006, pp. 573–580.

[6] S. Harabagiu and F. Lacatusu, "Topic themes for multi-document summarization," in *Proceedings of SIGIR 2005*, 2005, pp. 202–209.

[7] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization," in *Proceedings of NAACL/HLT 2009*, 2009, pp. 362–370.

[8] V. Nastase, "Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation," in *Proceedings of EMNLP 2008*, 2008, pp. 763–772.

[9] A. Nenkova and A. Louis, "Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization," in *Proceedings of ACL-HLT 2008*, 2008, pp. 825–833.

[10] F. Lacatusu, A. Hickl, K. Roberts, Y. Shi, J. Bensley, B. Rink, P. Wang, and L. Taylor, "LCC's gistexter at duc 2006: Multi-strategy multi-document summarization," in *Proceedings of DUC 2006*, 2006.

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

[12] B. Schiffman, I. Mani, and K. Concepcion, "Producing biographical summaries: Combining linguistic knowledge with corpus statistics," in *Proceedings EACL 2001*, 2001, pp. 450–457.

[13] H. Alani, S. Kim, D. Millard, M. Weal, W. Hall, P. Lewis, and N. Shadbolt, "Automatic Ontology-Based Knowledge Extraction from Web Documents," *IEEE Intelligent Systems*, pp. 14–21, 2003.

[14] E. Filatova and J. Prager, "Tell me what you do and I'll tell you what you are: Learning occupation-related activities for biographies," in *Proceedings of HLT-EMNLP 2005*, 2005, pp. 49–56.

[15] N. Garera and D. Yarowsky, "Structural, Transitive and Latent Models for Biographic Fact Extraction," in *Proceedings of EACL 2009*, 2009.

[16] F. Biadsy, J. Hirschberg, E. Filatova, and L. InforSense, "An Unsupervised Approach to Biography Production using Wikipedia," in *Proceedings of ACL-HLT 2008*, 2008, pp. 807–815.

[17] C. Sauper and R. Barzilay, "Automatically generating wikipedia articles: A structure-aware approach," in *Proc. of ACL 2009*, 2009, pp. 208–216.

[18] M. Pasca, "Turning Web Text and Search Queries into Factual Knowledge: Hierarchical Class Attribute Extraction," in *Proceedings of AAAI 2008*, 2008, pp. 1225–1230.

[19] H. Daume and E. Brill, "Web search intent induction via automatic query reformulation," in *Proceedings of HTL-NAACL 2004*, 2004, pp. 49–52.

[20] H. Zeng, Q. He, Z. Chen, W. Ma, and J. Ma, "Learning to cluster web search results," in *Proceedings of SIGIR 2004*, 2004, pp. 210–217.

[21] P. Cheng, C. Tsai, C. Hung, and L. Chien, "Query taxonomy generation for web search," in *Proceedings CIKM 2006*, 2006, pp. 862–863.

[22] Z. Zhuang and S. Cucerzan, "Re-ranking search results using query logs," in *Proceedings of CIKM 2006*, 2006, pp. 860–861.

[23] X. Wang and C. Zhai, "Learn from web search logs to organize search results," in *Proceedings of SIGIR '07*, 2007, pp. 87–94.

[24] Google Knol, http://knol.google.com/k.

[25] Yahoo! Glue, http://glue.yahoo.com/.

[26] "Lycos Retriever," http://www.lycos.com/retriever.html.

[27] S. Cucerzan, "Large-scale named entity disambiguation based on Wikipedia data," in *Proceedings of EMNLP-CoNLL 2007*, 2007, pp. 708–716.

[28] R. Mihalcea and A. Csomai, "Linking Documents to Encyclopedic Knowledge," in *Proceedings of CIKM 2007*, 2007, pp. 233–242.

[29] "Document Understanding Conferences. 2001–2007," http://www-nlpir.nist.gov/projects/duc/pubs.html.