

Rating News Documents for Similarity

Carolyn Watters*

Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia B3J 2X4, Canada.

E-mail: watters@cs.dal.ca

Hong Wang

Acadia University, Wolfville, Nova Scotia B0P 1X0, Canada

Electronic news has long held the promise of personalized and dynamic delivery of current event news items, particularly for web users. Although electronic versions of print news are now widely available, the personalization of that delivery has not yet been accomplished. In this paper, we present a methodology of associating news documents based on the extraction of feature phrases, where feature phrases identify dates, locations, people, and organizations. A news representation is created from these feature phrases to define news objects that can then be compared and ranked to find related news items. Unlike traditional information retrieval, we are much more interested in precision than recall. That is, the user would like to see one or more specifically related articles, rather than all somewhat related articles. The algorithm is designed to work interactively with the user using regular web browsers as the interface.

1. Introduction

In this paper, *news* is information about recent events of general interest, especially as currently reported by newspapers, periodicals, radio, and television. Gigantic archives and repositories of this news data (text, photos, audio, and video) are now being built in digital format from a variety of sources each day. This news data is often used in text analysis, querying, and text mining largely because of its volume and general interest.

We concentrate on the use of current, dynamic news data rather than archival data. This is an important distinction because the value of retrieved articles depends on the context of the user. Consequently, the approach to news delivery must relate to the context within which such a system would be used. What do people expect to gain from reading the news, where typical news sources are newspaper, radio, and television? Two behavioral theories apply to news reading: uses and gratification and play or ludenic. The uses

and gratification theoretical perspective is based on the assumption that the reader has some underlying goal, outside the reading itself, that reading the news satisfies. That is, "... an assumption that media use, including news reading, serves some ulterior purpose external to the communication behavior itself." (Dozier & Rice, 1984) This perspective implies that optimal content and form can be determined once the particular information goal is known. The uses and gratification theory applies when the user is accessing newspaper databases or clipping services as opposed to simply "reading the news." When accessing a newspaper database, the user is attempting to satisfy an explicit information need, and must be able to express this need in terms of a query or a profile description. Newspaper databases with online access are essentially document retrieval systems in which individual news items are treated as discrete units (i.e., news items are treated as documents) (for a review, see Berman, 1993; Pack, 1993). Such systems typically provide retrieval in response to a user query and/or personalized clipping services (selective dissemination of information) based on user profiles.

Task analysis of user interfaces and information systems are generally based on the theory of uses and gratification. That is, the user has some goal or task that can be satisfied by using the system and the evaluation of the system measures the degree to which that goal has been achieved.

The ludenic or play theory of news reading (Stephenson, 1967) asserts that, "... the process of news reading is intrinsically pleasurable, and that intrinsic pleasure is at the root of a mature, orderly, and highly ritualized form of news reading as well as a more casual, spontaneous, and unstructured form of news reading." (Dozier & Rice, 1984).

Several ludenic behavioral characteristics are consistent with news reading behavior: individual path selection (convergent selectivity), apperception, and habitualness. Individuals read the paper differently. They select different items to read, read articles in different orders, and read different amounts of items selected in the paper. Each

*To whom all correspondence should be addressed.

individual generates a unique path (convergently selects a path) through the news items and that path is pretty nearly impossible to predict (Allen, 1990; McGillivray, 1995). Furthermore, humans apply their capability for apperception to news reading. That is, readers perceive only those aspects of a complex situation that fits within their current interests and/or understanding. Thirdly, news reading behavior, as characterized by Stone and Wetherington (1979), is an habitual activity that, "... accompanies daily rituals and is typically performed in the same place at the same time; indeed late deliveries lead to cancellations."

The ludenic theory of news reading implies that news presentation systems must assist the users by making it easy to browse, easy to skim or to read in depth, by providing a comfortable sameness, and by making the process itself an enjoyable part of the day. "What ludenic newsreaders require is an *edited* product, shaped narrowly enough in form and content to permit convergent selective processes to occur *through* protocols that are pleasurable ends in themselves" (Dozier & Rice, 1984). Although electronic news sources must accommodate both goal-oriented and ludenic uses, reading the news is primarily a social activity, and users need access to a presentation mode that allows them to select items of interest and to enjoy the process. Success in "getting the news" is a very vague concept and seldom do any *a priori* queries exist against which results can be measured. Task analysis may not be appropriate to behavior that is not specifically goal oriented, and that has a large social component. The focus of this work is to examine how the system can assist the user in selecting items of interest to augment an edited selection presented to them. Such a system is not primarily concerned with retrieving all related items, but rather retrieving one or possibly a few articles that are specifically related to the current item.

2. Background

A number of research projects are working on electronic news delivery systems, into which the developments reported in this paper would fit. The Electronic News Delivery Project (Watters et al., 1998) at Dalhousie, Acadia, and Waterloo Universities have developed an overall architecture that integrates text, photographs, video, and audio into personalized multimedia news presentations. This architecture has three layers: resource layer, management layer, and reader layer. The resource layer stores the actual news items. The management layer decides the content of a given presentation. The reader layer configures the presentation for the reader.

Other projects have focused on selection and querying of news information such as Janne et al. (1998) and the INFOS (Intelligent News Filtering Organizational System) (Kendrick & Rao, 1998). Statistical methods and natural language processing have also been used to improve the effectiveness and efficiency of retrieval of news items (Yan & Garcia-Molina, 1995). The ANES (Automatic News Extraction System) and Searchable Lead (Rau, 1994) both use a

combination of statistical and heuristic methods to extract key sentences for summation. SCISOR (System for Conceptual Information Summarization, Organization and Retrieval) (Jacobs & Rau, 1991) performs text analysis and question answering on financial news using a combination of bottom-up and top-down discovery of linguistic structures. Carrick and Watters (1997) used a frame discovery method to automatically associate photos with related news items.

Agents have also been used to search news articles. NewsHound is a Knight Ridder service that uses the Verity search engine to search the contents of the San Jose Mercury News and other news sources to select articles that fit a user profile. Verity (*Business Wire*, 1999) is a commercially offered extended Boolean retrieval engine for searching news items using complex Boolean and proximity matching algorithms to identify "concepts." Typically, these agent driven systems have an indexing engine that binds key words or concepts to each article in the news set. These are then processed against a persistent user profile of interests to return a set of documents filtered by those expressed interests.

3. Architecture

The design is a three-tier web architecture: client, web server, and middle analysis tier. The client is very thin, relegated to browser display and user interaction. The web server handles web interaction, and the middle tier manages the data analysis and similarity analysis tasks. The middle tier in this architecture has two main functions. The first is to extract attribute values from individual news items to build news representation of those items. The second is to compare these news representations for similarity.

3.1. Defining the News Representation Object

Ideally, a representation of a news item should provide answers for the six classic news questions: who, what, where, when, why, and how. Of these six attributes, three can be extracted reasonably accurately and reasonably quickly: who, where, and when. The other three attributes, what, why, and how, require understanding of the content of the article, a task that is complex, computationally heavy and not yet well understood. Our goal is to provide reliable links very quickly from one news article to others, and so we are not interested in algorithms that are computationally intensive or that return uncertain results. The task in this scenario is to get related articles, not to get all relevant articles.

We define a representation of each news item as an object, where an instance of such an object has attributes and behaviors consistent for that object type, as shown in Figure 1. The attributes of the news object fall into two categories: header and content. The header attributes are structural and include: author/byline, publisher, title, date. The content attributes are location, date, name, and organi-

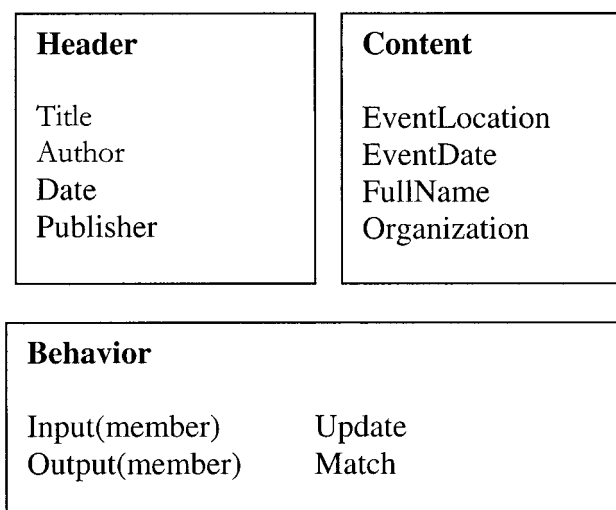


FIG. 1. A news object.

zation. Arrays of name-phrases can be associated with each content attribute for each instance of this object, where a name-phrase can be identified as a sequence of one or more words beginning with capital letters, for example, "Ottawa" or "Acadia University." Earlier work (Carrick & Watters, 1997) showed the effectiveness of using proper nouns and names to identify important news attributes.

Behaviors or methods can be defined for the news representation object that includes: input, output, update, and match. Next we can define the class of such objects as the News Class, shown in Figure 2.

3.2. Generating News Objects

There are three steps in the process of generating a news object from a news item in marked up form. The structure of the news objects was used to generate a common marked up form automatically. The first task is to extract the structural information, title, date, publisher, etc. The second task is to produce the set of potential name-phrases from the news document. The third task is to categorize these name-phrases into the content feature attributes (see Fig. 3). Figure 4 outlines the procedure, in general, for creating an instance of a news object from a marked up news article.

Extraction of Structural Information. Generally, news data either comes with some recognizable mark up included in the text or with identifiers for the common structural components. In our case, the *Halifax Herald* data is made available to us with an XML-like mark up in place. This makes it a straightforward process to instantiate attributes such as title, date, and author.

Extraction of Name-Phrases. In English, at least, capitalization generally indicates that the word or word phrase represents the name of a location, date, person, or organization. These name-phrases can then be used to instantiate content attributes of the news object. The first step is to

identify all single occurrences and sequences of capitalized words. Periods within sequences of capitalized words are examined to determine whether they identify an abbreviation, such as "Mr." or "Ltd." or the end of a sentence, as in "Spokane. Clinton." Other punctuation, such as commas, apostrophes, brackets, and dashes are also examined for context and may or may not delimit a name-phrase. Common conjunctions, such as "and" or "of," sandwiched between capitalized words, are also left in. Terms occurring at the beginning of each sentence are assumed initially to be proper nouns and only categorized as names or organizations following further analysis. A stop list of high-frequency noise words is used to reduce the size of this set for the document. The stop list was generated by combining a standard stop list (Salton, 1989) with a compilation of 343 most frequent capitalized words in the previous 150 editions of the *Halifax Herald*. Duplicated occurrences of name-phrases are also removed at this time. Table 1 shows the results of the extraction phase on the sample news article in Figure 3.

The algorithm then concentrates on the analysis of the phrases in this list to instantiate the attributes of the final representation, shown in Table 2, for the sample article. One can see that the information value of this list of terms is more focused than the key words, which would be extracted from this document and broader than the information value of the title plus first paragraph, which is often used for news.

Categorization of Name-Phrases. The next step is to categorize the name-phrases in the set into the four categories of interest: location, date, person name, and organization. Previous experience with news data (Carrick & Watters, 1997) showed that it is worth saving the leftover names for search purposes. In that study, almost every successful match of caption to news articles included at least one "other" attribute. In this study, we put the uncategorized names in the organization category for weighting purposes.

Event Location. A name-phrase is categorized as a location if it matches the name of a politically or geographically defined location (city, province, state, country, international region, body of water, mountain, etc.). A dictionary of 2444 geographical names from around the world from *Energy, Mines, and Resources Canada* was used. This dictionary database expands as the system is used and more geographical names are identified.

Event Date. A name-phrase is categorized as an event date if it contains the names of days or months, from a small dictionary of 37 variations.

Fullname. A fullname is the name of a person or family. We used a dictionary of 3190 common first names, a dictionary of 3200 last names accumulated from newspaper data, and a list of 33 common titles to identify name-phrases in this category. Fullnames may have several subcomponents that we categorize as: other, title, first name, middle name, and last name. For example the name-phrase "MT&T President John Martin J. Jones" would be identified as the following fullname attribute (Carrick & Watters, 1997).

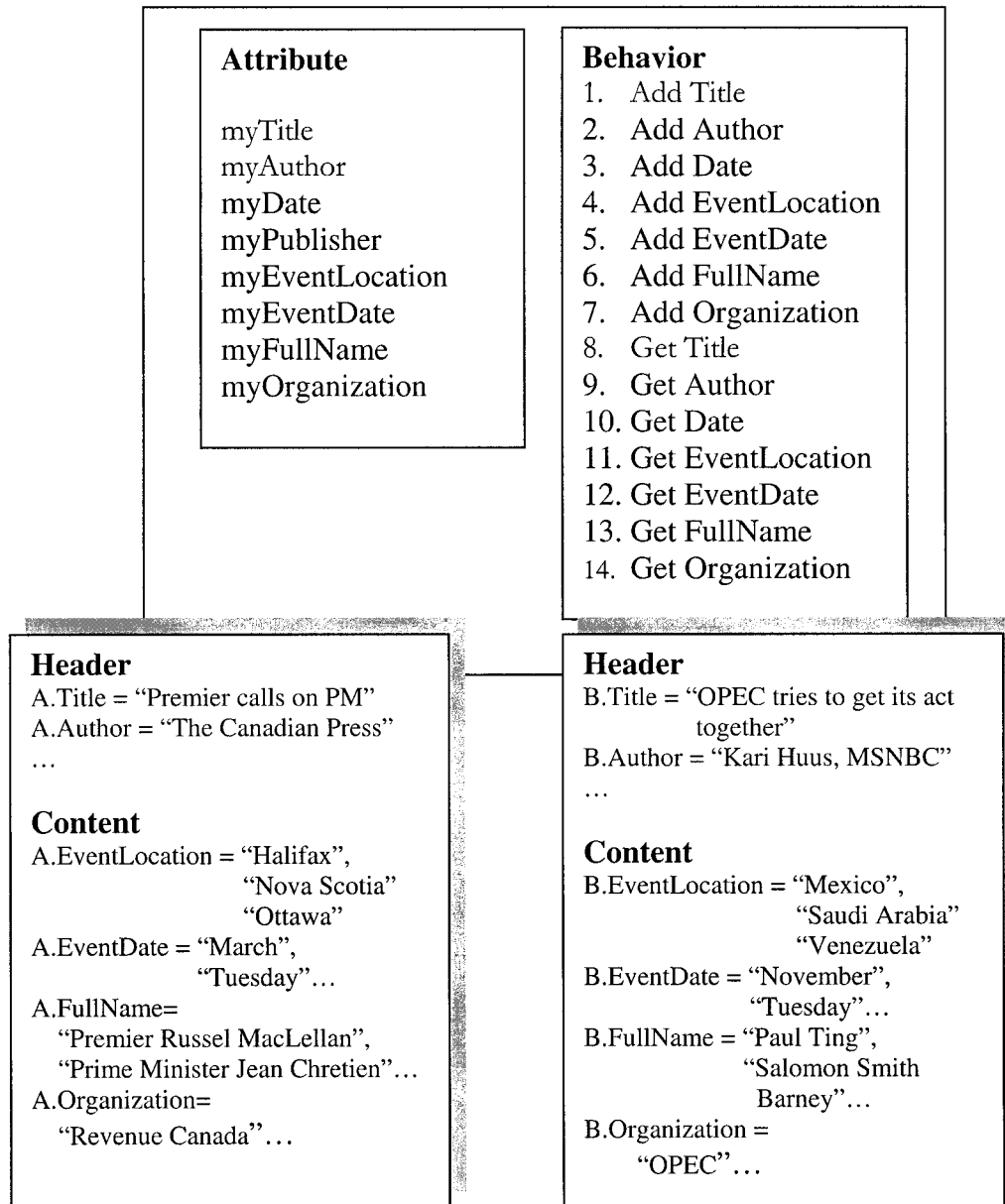


FIG. 2. A and B are instances of the class News.

FullName	
Other	MT&T
Title	President
FirstName	John
MiddleName	Martin
MiddleName	J.
LastName	Jones

To be a fullname, the phrase must have an identifiable last name. The last name may be determined either from a match in the last name dictionary or from any currently identified fullname attribute.

Searching on titles was a highly effective means to identify person's names in news articles tested. We are able to take advantage of the news reporting protocol of using the person's full name once, such as "President Bill Clinton," and then referring to that person subsequently by the last name, "Clinton," only.

One of the interesting underlying characteristics of current events is that the names are a continuum. Consequently, the name database is continually updated as new names are identified.

Organization. The organization category includes all of the name-phrases not yet categorized. The main interest in

```

<PUBDATE>
1998/03/01
</PUBDATE>
<HEADLINE>
Premier calls on PM
</HEADLINE>
<BYLINE>
By THE CANADIAN PRESS
</BYLINE>
<CONTENT>
Premier Russell MacLellan came to town Tuesday to drop in on some old friends - like Prime Minister Jean
Chretien and Finance Minister Paul Martin.

MacLellan, heading a precarious Liberal minority government, has been under pressure from opposition New
Democrats and Conservatives over the blended sales tax the province shares with Ottawa.

The premier has also been pressing the federal government for extra funding for post-secondary education to
take account of the large number of out-of-province students in Nova Scotia universities.

When spotted on Parliament Hill, however, he was close-mouthed about his visit. He did acknowledge he'd
"probably" see Chretien before catching an evening flight back to Halifax.

A spokesman for the prime minister confirmed a meeting was planned but said it was private and there would
be no further comment.

An aide to Martin similarly confirmed MacLellan would be meeting the finance minister but said there would
be "no formal agenda."

MacLellan and Martin have met before - once last October and once in February - to discuss the blended sales
tax.

Then came the March 24 election that reduced the Liberals to 19 seats, a flat-footed tie with the NDP, while
the Tories took 14.

Both parties have threatened to withhold support for MacLellan's government, and the budget he wants to
bring in this spring, unless there is movement on the tax front.

The NDP campaigned on a platform of scrapping the BST while the Conservatives want significant changes.

MacLellan promised, during his run for the Liberal leadership last year, to offer tax rebates for home heating
oil and electricity. But once in power he found the lost revenue would endanger another key promise - to
balance the provincial budget.

The premier said after his last meeting with Martin that there was no prospect for Ottawa directly funding any
BST relief.

But he did ask for help expediting a Revenue Canada decision on alleged tax overpayments by Nova Scotia
Power. The province is seeking refunds that would help ease the $16-million burden the BST added to
electricity bills.
</CONTENT>

```

FIG. 3. An example of a marked up file. From the March 1, 1998 *Halifax Herald*, The *Halifax Herald* is published by The Chronicle-Herald Ltd., Halifax, Nova Scotia, in both print and electronic forms. The *Herald* is the largest daily newspaper in the province.

this category is the organization name subcategory. Here, an organization refers to a corporate, government, or other organizational entity, such as university, department, or school. These names are very hard to identify, especially in the context of world events. Rather than lose these name-phrases as potential match components, we decided to risk

including noise phrases in this final category. Roughly half of the name-phrases end up in this category. Most of the "noise" phrases drift down to this category. This does not unduly influence matching of articles by organization phrases as the attribute lists are compared to each other (i.e., organization list against organization list only).

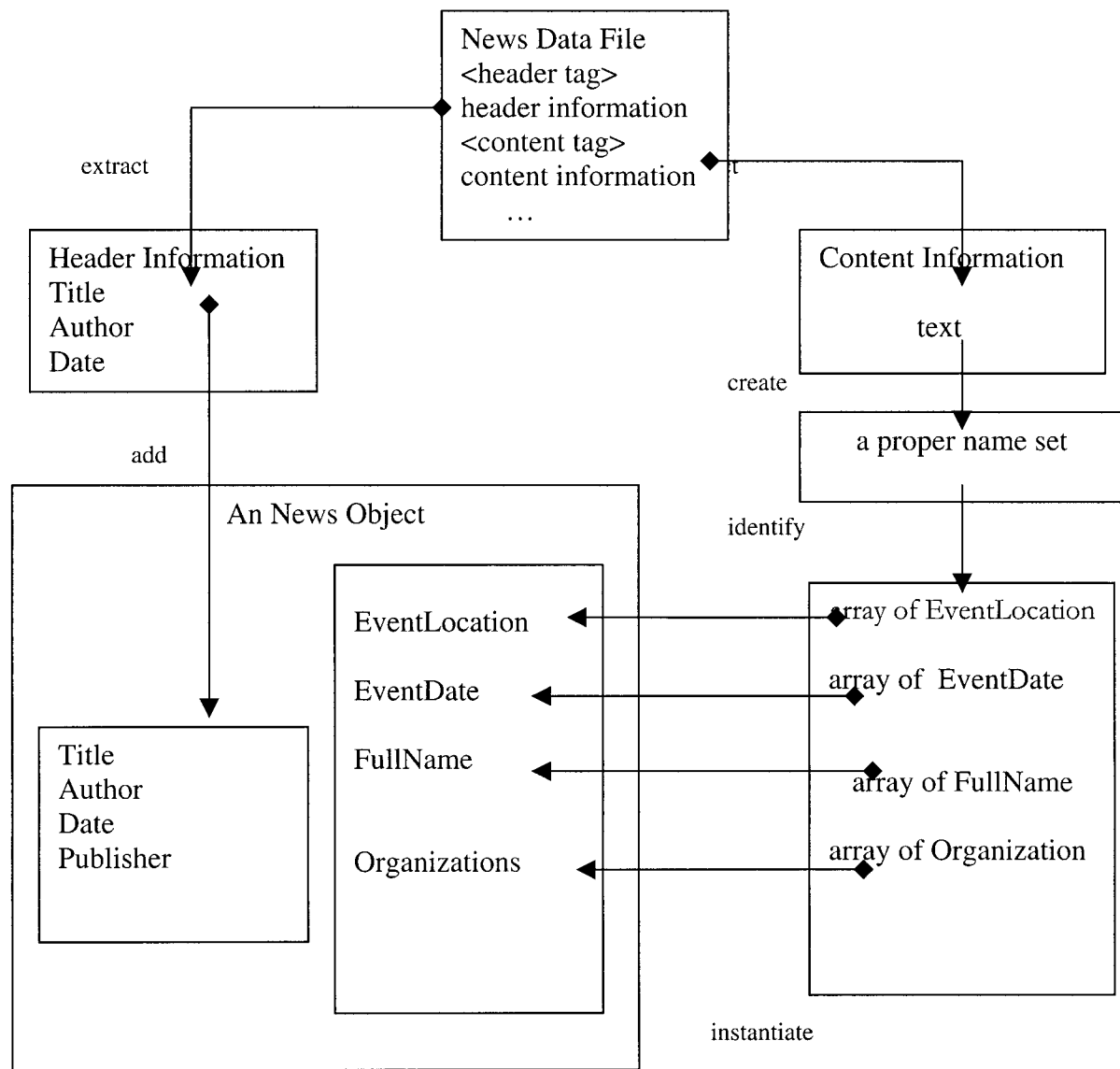


FIG. 4. The process of creating a news representation object.

Reduction of Redundant Name-Phrases in Categories. Within feature categories, obvious redundancies can be determined. For example, if both “Wed.” and “Wednesday”

TABLE 1. The proper name set (using the document presented in Figure 3 as input data).

Premier Russell MacLellan	Tuesday
Prime Minister Jean Chretien	Finance Minister Paul Martin
MacLellan	Liberal
New Democrats	Conservatives
Ottawa	Nova Scotia
Parliament Hill	Chretien
Halifax	Martin
October	February
March	Liberals
NDP	Tories
BST	Revenue Canada
Nova Scotia Power	

are in the eventdate category array, then only one is kept for the representation object. This process is relatively simple for the eventdate and eventlocation categories and less simple for the fullname and other categories.

In the case of fullnames, a partial match process is used to determine redundant phrases. In general, if two fullname name-phrases have the same values for the first name and last name, the shorter one is eliminated. So, for example, if both “Ms. Sally Jones” and “Sally Jones” occur in the article, then the shorter phrase “Sally Jones” is eliminated. If a fullname consisting of only the last name has the same last name as one or more other name-phrases, then it is eliminated. So, for example, if “Sally Jones” and “Jones” occur within the same article then the shorter phrase, “Jones” is eliminated.

For the name-phrases in the others category, a simpler match process is used to reduce redundancy. If a name-

TABLE 2. The feature name-phrase set (using the document shown in Figure 1 as input data).

Title	Premier calls on PM
Author	By THE CANADIAN PRESS
Date	1998/04/01
Event location	Halifax Nova Scotia Ottawa
Event date	March February October Tuesday
Full names	[O] Premier [T] [F] Russell [M] [L] MacLellan [O] Prime Minister [T] [F] Jean [M] [L] Chretien [O] Finance Minister [T] [F] Paul [M] [L] Martin
Organizations	Parliament Hill Conservatives Liberals NDP Tories BST Revenue Canada Nova Scotia Power

phrase is part of another one it is deleted. So, for example, given both “Respite Care Unit” and “Victoria General’s Respite Care Unit,” “Respite Care Unit” is contained in the second name-phrase and would be eliminated.

It is interesting to note the proportions of phrases within the categories. Both Carrick and Watters (1997) and this current study indicate that roughly 50% of the name-phrases are organization/other names, with people names accounting for roughly 25% of the phrases identified. More than 90% of the person names appearing in the news articles used for the trials were correctly identified.

3.3. Document-Document Similarity Measure

The goal of the news representation is to facilitate selection of other news items similar in some specified way to a current item. The user may opt for automatic similarity based selection of other articles, or the user may interact and select the features of interest to drive the selection process. For example, the user may want items with the same people and/or same event and/or same organization. In this work, we used a similarity measure to measure the similarity of two news items, where both items have been processed and the name-phrases used to instantiate attributes. Although standard matching functions such as the Cosine or Jaccard functions could have been used, we chose the simple function described below to accommodate several divergences from the vector model. The similarity of news items is dependent on the similarity of attribute lists rather than the

similarity of attributes within the lists, the attribute lists are of indeterminate length, and the importance of a given attribute is dependent, to a certain extent, upon the length of the attribute list.

A similarity measure between 0 and 1 is calculated between pairs of news representations with a threshold that can be adjusted by the user but had a default value of 0.40.

The basic similarity between two news representations is calculated using

$$\text{sim}(A, B) = \alpha \sum_{i=1}^4 a_i$$

where a_i is the similarity of the i th feature in the news representation for documents A and B and where α is a scaling factor, calculated by

$$\alpha = \frac{1}{\sum_{i=1}^4 \min(|A_i|, |B_i|)}$$

and $\min(A, B)$ is the length of the smaller of the two name-phrase sets for that feature. The value a_i is calculated then as

$$a_i(X, Y) = \sum_{j=1}^m \sum_{k=1}^n \theta(x_j, y_k)$$

which is the sum of common terms between the name-phrase sets for X and Y . m is the length of X , and n is the length of Y . The value of θ is calculated by

$$\theta(x_j, y_k) = \begin{cases} 0 & x_j \neq y_k \\ 1 & x_j = y_k \end{cases}$$

In the case of dates and locations, x_j and y_k are only considered to be the same when they precisely match. For example, May 1999 and May 6 1999 are not deemed to be same nor are “Beijing China” and “China.” In the case of organizations, however, x_j and y_k are considered to match when one is a substring of the other. So, “Beijing KL Technology Co” is deemed to match “KL Technology Co.”

Full names require special rules for calculating the similarity between name-phrases. Each fullname term has five attributes that contribute to the value, and each has an importance weight associated with it. A minimum requirement is match on the last name. These importance weights and a threshold value were determined by trial and error and are very close to those used by the earlier study (Carrick & Watters, 1997).

Attribute	Importance weight
Title	0.20
First name	0.30
Middle name	0.05
Last name	0.40
Other	0.05
Threshold	0.80

The following equation is used to calculate the similarity of two fullname-phrases:

$$\text{Sam}(N_1, N_2) = \sum_{i=1}^5 w_i \theta(n_{1i}, n_{2i})$$

where w_i is the importance weight of the i th attribute. $\theta(n_{1i}, n_{2i})$ is calculated using

$$\theta(n_{1i}, n_{2i}) = \begin{cases} 1 & n_{1i} = n_{2i} \\ 0 & n_{1i} \neq n_{2i} \end{cases}$$

The following three examples show the outcome of this calculation.

Name 1	Name 2	Result
John Smith	Mr. Smith	1.0
Mr. Smith	Justice Smith	0.8
John Smith	George Smith	0.7

Recall that all matches must have a precise match on last name. Given that, title difference or title without a first name are not enough to distinguish between two individuals. A difference in first name, however, does, generally, distinguish individuals.

4. Prototype

4.1. Sample Session

The user begins with a typical newspaper broadsheet arrangement of articles from one or more news sources. In Figure 5, the news items are from the *Halifax Herald* of that day with Yahoo news additional items. This works like any other online “newspaper.” The user clicks on a story link to get the full article or on a section link to get the sports or world news, etc. The filtering tools provide the user with direct access to the dictionaries and news object representations. This provides the user access to backend functions and tools for updating dictionaries, generating attribute lists, or adjusting thresholds (see Fig. 6).

The user now has the option *related* that retrieves documents from the news collection based on name-phrase similarity to this article. In our sample dataset, the news set includes 25 world news articles from Yahoo, MSNBC, and Reuters. The user can direct the similarity match process by selecting name-phrases of interest on which to base the

similarity process using a screen like the one shown in Figure 7. The user can set a limit on the number of similar articles retrieved and may adjust the similarity threshold. In our trial sessions, we did not impose a limit on the number of articles retrieved, as the dataset was not large.

In this case, the user has requested that location should be London, the people Canadian Louise Frechette and President Bill Clinton, and events of interest are UN General Assembly and Security Council. In this case, the only other article with a similarity measure more than the threshold is Israel Adopts UN Lebanon Pullout Decision with a similarity of 0.6. The result of this is a new page of hot links to those articles with similarity more than the threshold.

If the user does not select any name-phrases, then all of the name-phrases are used in the similarity measurement algorithm.

4.2. Sample Results

Informal trials were run using 78 online news items from the daily edition of the *Halifax Herald* to examine the performance of the dynamic generation of news representations and subsequent application of the similarity algorithms to retrieve related news items extracted from other news sources. In one trial, we examined the accuracy and efficiency of the algorithms and the second trial evaluated the usefulness of the results from the user perspective. A sample set of 25 news items from Reuters, AP, MSNBC, and Yahoo!Asian sites were included to form a domain of “other items.” In this set, were articles on Kosovo, Iraq, and Asia.

The test set of 78 news items from April 1, 1998 *Halifax Herald* news service consisted of 28 local items, 15 national items, 14 international items, 10 business items, and 11 entertainment items. This dataset showed had the following characteristics:

Category	Total words	Capital terms	Important terms
Local news	8376	1315	1067
National news	6075	748	575
International	4533	612	472
Business	3857	607	498
Entertainment	3477	557	439

The important terms were identified by manual inspection. These were terms the human reader identified as “important” and were used to see how well the algorithm generating the capital term list performed. Overall, the frequency of capital terms selected from articles by the algorithm was fairly consistent over the range of categories, ranging from 12.3% for the national news articles to 15.7% in the local news articles. The proportion of the capital terms deemed to be important terms is quite high, ranging between 77–82% in the sample. This means that we can use relatively simple processing to generate the capital term list

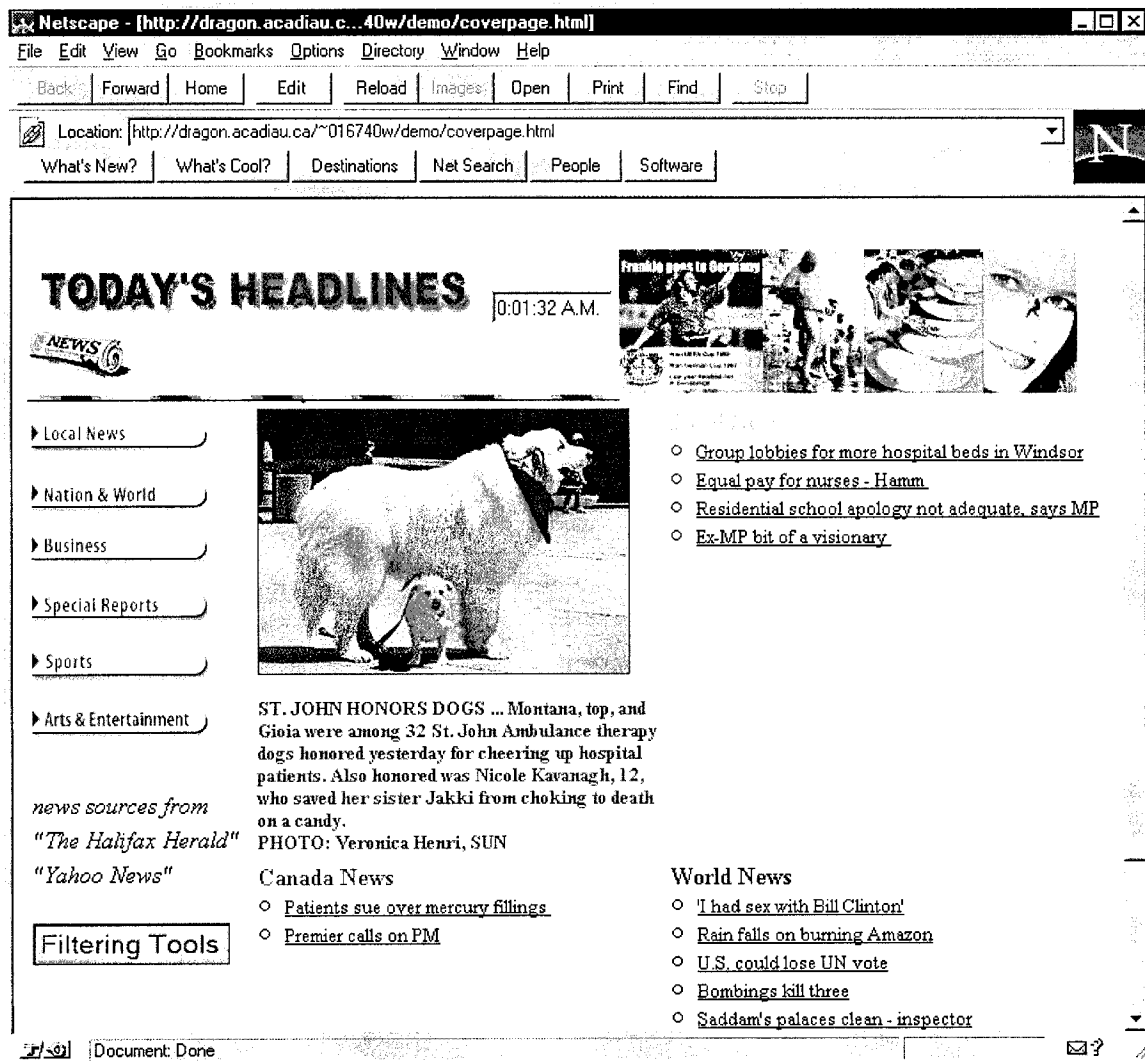


FIG. 5. Sample opening page.

and concentrate our computationally heavy algorithms on this list.

The performance metric of importance in this instance is the accuracy of the feature extraction from the capital term list. If a term is incorrectly assigned to be a location rather than a person's name, then later similarity measures are bound to fail. Unlike bibliographic retrieval performance, the user expects all related items to be very accurate matches. The accuracy of selection of dates and location terms was very high in the base set of articles, nearly 100%, whereas the accuracy of fullname term extraction was less at 93% for the same data. The accuracy of selection of the important phrases from the news articles was done by manually comparing the human selected terms against the list of capital terms selected by the algorithm. The accuracy was fairly consistent at about 90%. That is, for every ten important phrases identified by a person in these articles, the algorithm found, on average, nine.

The correct identification of person names was over 90% whereas the correct identification of locations and dates was

100% in the articles tested. The precision of the document similarity algorithm in three trials from a base set of 78 articles from the *Halifax Herald* plus 25 articles from Yahoo news, the precision of related stories was very promising, with 100% precision at a threshold of 0.40. The recall level at this threshold level was, however (and predictably), lower at 100%, 93%, and 45%. Our concern in this system is not recall and so no further reductions in threshold were tested. The user decides which name-phrases to use in the search for related articles. We found that the best results occur when the user includes person names in the match list.

	Local	national	internat	business	Entertain
Identified phrases	344	200	166	241	130
Found phrases	317	182	149	212	116
Association rate	92%	91%	90%	88%	89%

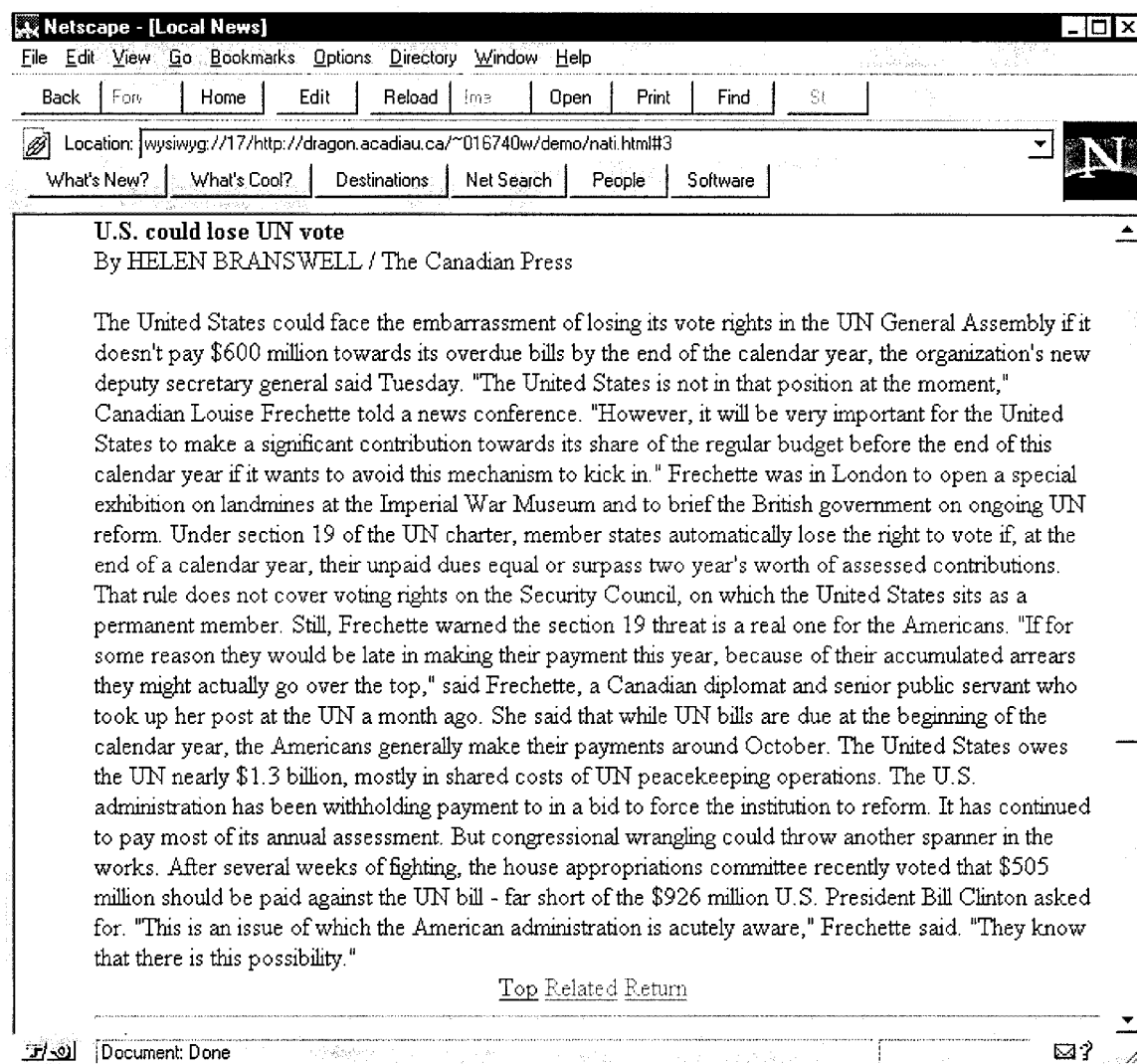


FIG. 6. Sample article.

We found two areas caused most of the difficulties leading to missed or incorrect feature extraction: mixed capitalizations and punctuation. Occasionally, features would have components that had embedded terms not starting with a capital letter, such as "Lac du Ronge," or "John and Sons," which were mistaken as two separate phrases, which are not particularly useful. If the punctuation is incorrect, which is often the case in news articles, then the algorithm is thrown off. In both cases, adding complexity in the feature extraction algorithm did not result in significant improvement of accuracy.

Two factors can be isolated as contributing to the usefulness of this system from the user perspective: precision of results and ease of use. In this system, users select news items from their "newspaper" and ask for other related items. The requirement for recall is secondary to the requirement for precision in this scenario.

Of the six users we had use the sample data, recall varied widely, from 45% to 100% whereas precision remained stable from 93% to 100%. In general, the selection of

name-phrases by the user affects the results more than other features. This is at least partly because the name-phrases tend to have fewer noise words among them.

The effect is to categorize, filter, and rank key word results in the same dataset.

4.3. Scale up Issues


Systems of this nature often have a scale up problem. The test datasets we used were in the hundred-article range, and not the millions of documents available on the web. We must consider whether the algorithm is applicable to larger datasets and more users. There are several indicators that this approach for this task domain is suitable for use in the "real world."

The size of the news domain on the Internet is, in fact, relatively small and well contained. Although there are at least 1000 newspapers now online in the United States alone, the reliance on wire service for much of the content

Netscape - [World News Rep Page]

File Edit View Go Bookmarks Options Directory Window Help

Back Forward Home Edit Reload Images Open Print Find Stop

Location: 

What's New? What's Cool? Destinations Net Search People Software

Title U.S. could lose UN vote
Author By HELEN BRANSWELL / The Canadian Press
Report Date 1998/04/01

Event Date	Event Location	FullName	Others
Oct	U.S.	Canadian Louise Frechette	UN General Assembly
Tue	British	President Bill Clinton	Frechette
	London		Imperial War Museum
			Security Council
			Americans

EventDate EventLocation FullName Others

[Story Top Return](#)


 Document: Done

FIG. 7. Sample similarity interaction.

means that a relatively few articles are replicated many times. The typical news edition has 100–200 articles and photos, many of which are from one of the wire services. Furthermore, the requirement for local news coverage is restricted to relatively small numbers of users and of little interest to others. This means that 100–200 articles is a reasonable news domain for most readers of general news (i.e., not topic researchers).

It is important to remember that such a system is not meant to replace search engines on the web but rather to work within some restricted domain of news articles, possibly a selected set of news servers. Each such news edition has several characteristics that help restrict the domain of possible articles. Almost all editions of news have relatively few news articles (plus a lot of advertisements), typically 100–200 items, with only three or four per page. Second, almost all editions of news have a standard categorization scheme, such as world news, national news, local news, sports, business, etc. This broadly used categorization can be used to determine the most likely subset of articles within which to search. Third, the number of unique sources of news articles is very small: Reuters, AP, etc. The exact same items are routinely printed in almost all news editions.

Although speed was not the main consideration in this work, the algorithm itself is fast enough to process documents in real time. Significantly, only 10–15% of the text is used in the analysis as 85–90% of the words are noncapitalized. Furthermore, no documents need to be processed more than once. So frequent use by a user or a large group of users does not require reprocessing the same articles over and over. The algorithm is on a need-to-use basis and so only executes as needed and, consequently, not all documents need to be processed for any given session.

The goal of the process is not to find all possibly related items, rather just more. This means that the algorithm does not process the dataset exhaustively rather selectively. Consequently, we argue that such an algorithm will scale up both to a handle a realistic news domain of items and a large number of users.

5. Summary

The system described uses similarity measurements to find related items based on interaction from the user and in real time. At the same time, it provides an efficient way to classify and group news articles on the fly. The name-phrase

filtering algorithm works well for news articles that typically deal with events, places, organizations, and people. These artefacts make good name-phrases. Also, the typical news reader is not interested as much in complete and exhaustive coverage of a topic so much as another article. So precision is much more important in this situation than is recall.

The main weakness in this approach is the effectiveness of the name-phrase extraction process. It was found that slightly different algorithms were needed to process different types of news articles. For example, the algorithm for business stories has to cope with numbers and company names whereas the algorithm for sports stories can be re-fined for scores, league names, and team names.

The main benefit of this approach is to provide related articles allowing input from the user based on a current article. Such an approach does not replace key word approaches for high recall or wide domain Internet searches, but provides better results in restricted news domains.

Acknowledgments

This work was partially funded by NSERC, the Natural Sciences and Engineering Research Council of Canada. The authors are grateful to the *Chronicle-Herald, Ltd.* for providing continuous access to the newspaper data.

References

Allen, R. (1990). User models: Theory, method, and practice. *International Journal of Man-Machine Studies*, 3, 511–543.

- Berman, M.A. (1993). Today's world news—creating a desktop news delivery system. In *Proceedings of the 14th National Online Meeting* (pp. 33–38).
- Business Wire* (1999). The globe and mail web sites deploy verity for comprehensive knowledge retrieval on Canadian national portal site. Press Release, April 26.
- Carrick, C., & Watters, C.R. (1997). Automatic association of news item. *Information Processing and Management*, 33(5), 615–632.
- Dozier, D., & Rice, R. (1984). Rival theories of electronic news reading. In R. Rice (Ed.), *The new media* (pp. 103–128). London: Sage Publications.
- Jacobs, P.S., & Rau, L.F. (1990). SCISOR: Extracting information from online news. *Communications of the ACM*, 33, 88–97.
- Kenrick, J.M., & Rao, V. (1997). Information filtering via hill climbing, wordnet, and index pattern. *Information Proc and Management* 33(5), 633–644.
- Korkea-aho, M., and Sulonen, R. (1997). Logical structure of a hypermedia newspaper. *Information Proc and Management*. 33(5), 599–614.
- McGillivray, K. (1995). Adaptive prediction of news items. Honours Thesis, Dalhousie University, Halifax, Nova Scotia.
- Pack, P. (1993). Electronic newspapers—the state of the art. In *Proceedings of the 14th National Online Meeting* (pp. 331–335).
- Rau, L.F. (1991). Extraction company names from text. *Seventh IEEE Conference on Artificial Intelligence Applications* (pp.29–32).
- Rau, L.F. (1994). Domain-independent summarization of news. Available: <http://www.cis.upenn.edu/~cliff-group/94/lrau.html#ref-lfr:summarization>
- Salton, G. (1989). *Automatic text processing*. MA: Addison Wesley.
- Stephenson, W. (1967). *The play theory of mass communication*. Chicago: University of Chicago Press.
- Stone, G., & Wetherington, R., Jr. (1979). Confirming the newspaper reading habit. *Journalism Quarterly*. 54, 554–561.
- Watters, C.R., Shepherd, M.A., & Burkowski, F.J. (1998). Electric news delivery project. *Journal of the American Society for Information Science*, 49(2), 134–150.
- Yan, T.W., & Garcia-Molina, H. (1995). SIFT—A tool for wide-area information dissemination. In *Proceedings of the 1995 USENIX Technical Conference* (pp. 177–186).