LocalSavvy: Aggregating Local Points of View about News Issues

Jiahui Liu and Larry Birnbaum

Northwestern University
Intelligent Information Laboratory
2133 Sheridan Road
Evanston, Illinois 60208 USA

j-liu2@northwestern.edu, birnbaum@cs.northwestern.edu

ABSTRACT

The web has become an important medium for news delivery and consumption. Fresh content about a variety of topics, events, and places is constantly being created and published on the web by news agencies around the world. As intuitively understood by readers, and studied in journalism, news articles produced by different social groups present different attitudes towards and interpretations of the same news issues. In this paper, we propose a new paradigm for aggregating news articles according to the local news sources associated with the stakeholders of the news issues. This new paradigm provides users the capability to aggregate and browse various local points of view about the news issues in which they are interested. We implement this paradigm in a system called LocalSavvy. LocalSavvy analyzes the news articles provided by users, using knowledge about locations automatically acquired from the web. Based on the analysis of the news issue, the system finds and aggregates local news articles published by official and unofficial news sources associated with the stakeholders. Moreover, opinions from those local social groups are extracted from the retrieved results, presented in the summaries and highlighted in the news web pages. We evaluate LocalSavvy with a user study. The quantitative and qualitative analysis shows that news articles aggregated by LocalSavvy present relevant and distinct local opinions, which can be clearly perceived by the subjects.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process* H.4.m [**Information System Applications**]: *miscellaneous*

General Terms

Human Factors, Design, Algorithms, Experimentation.

Keywords

News aggregation, local points of view, news source.

1. INTRODUCTION

As people increasingly turn to the web for information, online news reading is becoming more popular. Many news agencies are publishing and delivering news articles via the web. The wide variety of news content available on the web provides readers unparalleled opportunities for accessing news from around the world.

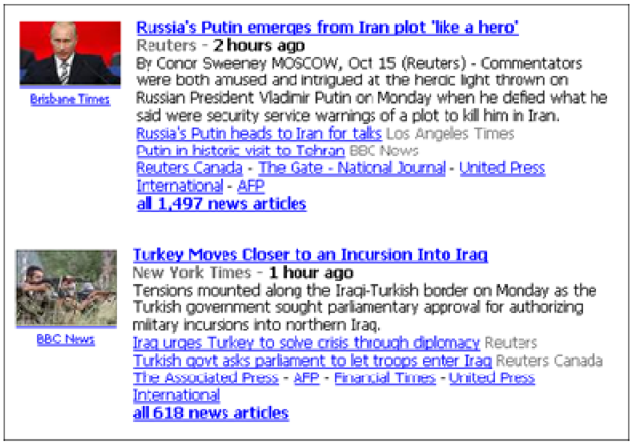
Typical news websites organize the large volume of news articles according to topics. Figure 1 shows a snapshot of Google News [12]. As is evident in figure 1, there is a huge number of news articles related to the same news issue. For example, there are 1,497 news articles about the event of Putin's visit to Iran. More importantly, the various news articles are published by sources located in different places. As discussed by van Dijk [26], news production is a subjective process affected by cultural norms and values. Journalists from different social groups may have different attitudes towards and interpretations of the same news event. For example, figure 2 and 3 illustrate snippets of news articles about Putin's visit to Iran from an American news source and an Iranian news source respectively. The two news articles present radically different points of view about the event, which are explicitly expressed in quoted opinions and implicitly expressed through different choice of words by the authors.

According to van Dijk [26], news articles from different social groups may be dissonant with the ideology held by readers, resulting in more interesting and memorable news reading. Finding the local views about issues in which readers are interested is also useful, for individuals, organizations and governments alike. For individuals, knowledge about other social groups can be illuminating and help them discover and overcome prejudices [21]. For policy makers of organizations and governments, the opinions of their counterparts are helpful for making informed decisions.

With the advance of the web as a popular medium for news delivery and consumption, news articles from a wide variety of places, organizations and other entities are more easily available than ever before. The wide availability of information on the web makes it possible for readers to gain local perspectives and insights into news events and topics. However, presenting news articles from every news source can be overwhelming to users. To gain leverage of this problem, we note that the *stakeholders*, the entities involved in or potentially affected by the issues, are most likely to express the strongest opinions, which are also most interesting to the readers.

In this paper, we propose a new paradigm for aggregating news articles according to the local news sources associated with the stakeholders of news issues. By aggregating news articles from sources local to the stakeholders of the news issue at hand, we can answer questions such as "Do the Iraqi people want U.S. troops in their soil?" or "How do the Chinese regard the human rights record of the United States?"

WWW 2008, April 21--25, 2008, Beijing, China.



Russia's Putin emerges from Iran plot 'like a hero'
Reuters - 2 hours ago
By Conor Sweeney MOSCOW, Oct 15 (Reuters) - Commentators were both amused and intrigued at the heroic light thrown on Russian President Vladimir Putin on Monday when he defied what he said were security service warnings of a plot to kill him in Iran.
Russia's Putin heads to Iran for talks Los Angeles Times
Putin in historic visit to Tehran BBC News
Reuters Canada - The Gate - National Journal - United Press International - AFP
all 1,497 news articles

Brisbane Times

Turkey Moves Closer to an Incursion Into Iraq
New York Times - 1 hour ago
Tensions mounted along the Iraqi-Turkish border on Monday as the Turkish government sought parliamentary approval for authorizing military incursions into northern Iraq.
Iraq urges Turkey to solve crisis through diplomacy Reuters
Turkish govt asks parliament to let troops enter Iraq Reuters Canada
The Associated Press - AFP - Financial Times - United Press International
all 618 news articles

BBC News

Figure 1. Snapshot of Google News

> Iranian President Ahmadinejad enjoyed a rare moment of superpower recognition this week with the arrival of Vladimir Putin to Teheran on the first visit of a Russian head of state to this country since 1943 … This new "Axis" alliance between Moscow and Tehran will leave the Security Council split and thus ineffective, forcing the United States- perhaps with Israeli support- to deal with the nuclear threat that it perceives as critical…
>
> *From U.S. Politics Today, DC*

Figure 2. News article published in U.S.

> TEHRAN, Oct 17--Leader of the Islamic Revolution Ayatollah Ali Khamenei has urged promotion of Iran-Russia ties, stressing Tehran's determination to satisfy its need for nuclear energy… "The Americans have repeatedly threatened the Islamic Republic since the Islamic Revolution, and they have spared no efforts to blow the Iranian nation, but their adventurisms have had negative effects," Ayatollah Khamenei said…
>
> *From Alalam News Network, Iran*

Figure 3. News article published in Iran

We implement the proposed paradigm in a prototype system called *LocalSavvy*. LocalSavvy analyzes the news web page that a user is reading. For the news event or topic described in the news article, LocalSavvy automatically identifies the people, organizations and places involved in or affected by the news issue as the *stakeholders*. The system then finds the locations with which that the stakeholders are associated and formulates queries to news search engines to gather articles from the appropriate local news sources. LocalSavvy retrieves the news web pages and extracts the local opinions from the pages. To assist effective browsing and reading, the local opinions are summarized in the aggregated results and also highlighted in the news web pages presented to the users.

LocalSavvy requires location related commonsense knowledge for analysis and searching, such as the country that a state official represents. Instead of using a pre-built static knowledge base, LocalSavvy uses the web as its knowledge source and dynamically gathers the knowledge via search engines. Its dynamic approach to knowledge acquisition alleviates the effort of knowledge engineering and keeps the system's knowledge up-to-date with the web.

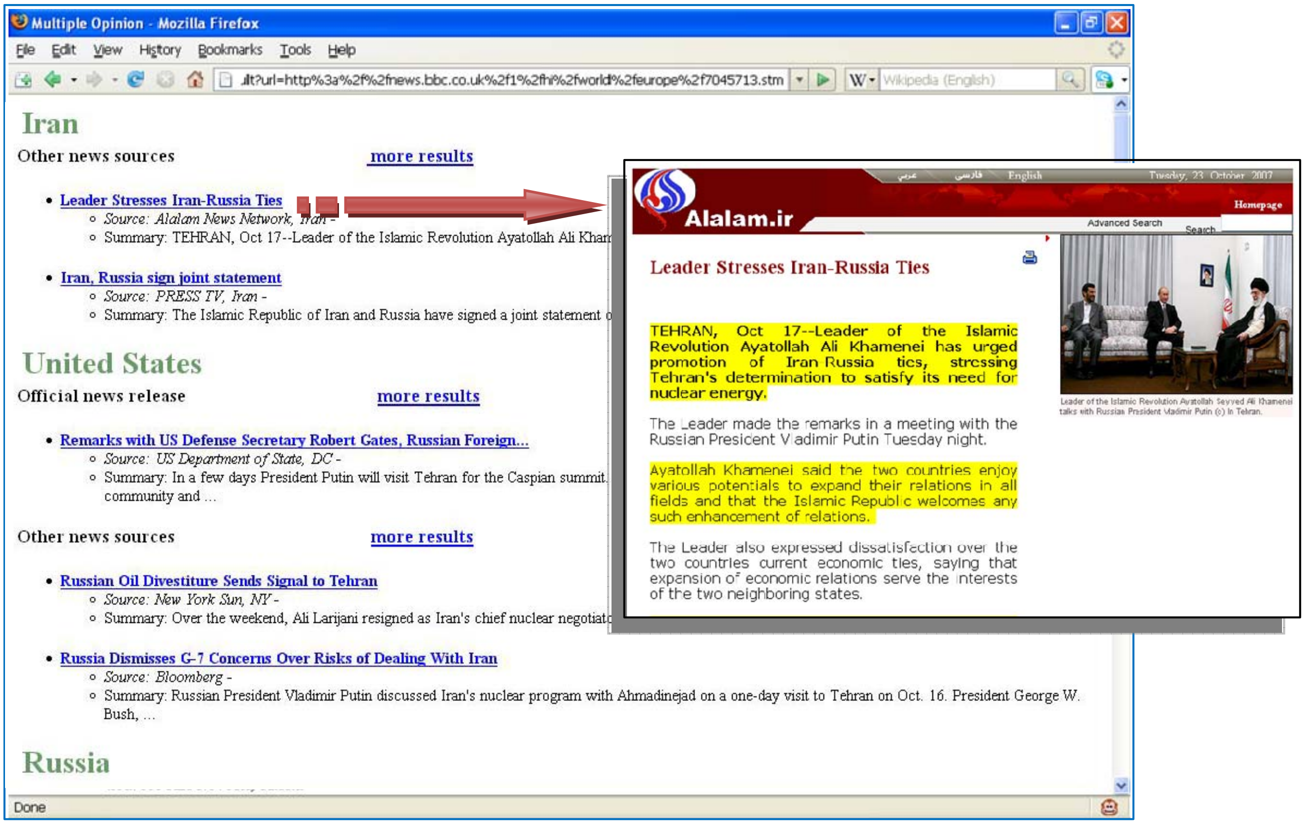
We conducted a user study to evaluate the efficacy of the system quantitatively with several metrics and qualitatively with questionnaires. The user study shows that the news articles published at different locations do present interestingly different points of view. Moreover, the news articles aggregated by LocalSavvy present the relevant and distinct local opinions, which can be clearly perceived by the subjects.

2. RELATED WORK

The paradigm proposed in this paper aggregates local points of view about news issues according to the location associated with the stakeholders. Some previous research has explored location information in web content. SmugMug [25] provides a service that allows users to map where they took their pictures on Google Earth. Christel et al. [8] investigated integrating spatial mapping with video library. GeoTracker developed by Chen et al. [6] aggregates RSS feeds according to the time and location in the content. Our study explores location information for a different purpose. Instead of finding about "What happened in X?", LocalSavvy aggregates news articles from the related places to answer "What do people in X think about it?". Moreover, LocalSavvy makes intelligent inference about the locations associated with the other entities mentioned in the news articles, in addition to the existing location information in the content.

LocalSavvy analyzes the news web page that a user is reading and aggregates related local views about the news issue it addresses. In this regard, it is closely related to research on context-based information retrieval. Previous research has studied modeling user's task-related documents to automatically retrieve useful information for the user [4, 5, 11, 19, 23]. Watson [4, 5], for example, analyzes opened text documents, such as web pages that the user is browsing and word documents that the user is authoring, formulates queries to web search engines and databases, and provides just-in-time information to the user. Systems like Watson identify relevant web pages based mainly on document similarity [4, 5, 11, 23]. Liu et al. [19] proposed that useful documents should be similar to the user's current context (or documents) in certain aspects, but systematically different in some other meaningful aspects. Compare&Contrast developed by Liu et al. [19] finds news stories about similar situations but involving different entities to support situation analysis. Our study explores another dimension of document difference in the context of news reading. By searching for news articles published in different locations, our system presents the various local points of view about the news issues.

There has been much research with regard to finding subjective expressions in documents. At the document level, Wiebe et al. [29] studied the identification of opinionated documents, such as editorials, using collocational clues of subjectivity in text. At the sentence level, Riloff and Wiebe [24] extracts subjective sentences from text using linguistic extraction patterns learned in a bootstrapping process. Ku, et. al [15] extracts and summarizes opinions in news and blogs based on a sentiment dictionary. More sophisticated problems, such as identifying opinion holders and topics, have also been investigated [7, 13, 14]. Our opinion extraction is distinctive in the way that it only extracts local opinions, i.e. opinions from the entities collocated with the stakeholders. To our knowledge, little research has been conducted on this problem.


Multiple Opinion - Mozilla Firefox
File Edit View History Bookmarks Tools Help
.ilt?url=http%3a%2f%2fnews.bbc.co.uk%2f1%2fhi%2fworld%2feurope%2f7045713.stm
Wikipedia (English)
عربي
فارسی
English
Tuesday, 23 October 2007
Alalam.ir
Homepage
Advanced Search
Search
Done

Iran

Other news sources more results

- Leader Stresses Iran-Russia Ties
  - *Source: Alalam News Network, Iran -*
  - Summary: TEHRAN, Oct 17--Leader of the Islamic Revolution Ayatollah Ali Kham
- Iran, Russia sign joint statement
  - *Source: PRESS TV, Iran -*
  - Summary: The Islamic Republic of Iran and Russia have signed a joint statement o

United States

Official news release more results

- Remarks with US Defense Secretary Robert Gates, Russian Foreign...
  - *Source: US Department of State, DC -*
  - Summary: In a few days President Putin will visit Tehran for the Caspian summit. community and ...

Other news sources more results

- Russian Oil Divestiture Sends Signal to Tehran
  - *Source: New York Sun, NY -*
  - Summary: Over the weekend, Ali Larijani resigned as Iran's chief nuclear negotiato
- Russia Dismisses G-7 Concerns Over Risks of Dealing With Iran
  - *Source: Bloomberg -*
  - Summary: Russian President Vladimir Putin discussed Iran's nuclear program with Ahmadinejad on a one-day visit to Tehran on Oct. 16. President George W. Bush, ...

Russia

Leader Stresses Iran-Russia Ties

TEHRAN, Oct 17--Leader of the Islamic Revolution Ayatollah Ali Khamenei has urged promotion of Iran Russia tics, strcssing Tehran's determination to satisfy its need for nuclear energy.

The Leader made the remarks in a meeting with the Russian President Vladimir Putin Tuesday night.

Ayatollah Khamenei said the two countries enjoy various potentials to expand their relations in all fields and that the Islamic Republic welcomes any such enhancement of relations.

The Leader also expressed dissatisfaction over the two countries current economic ties, saying that expansion of economic relations serve the interests of the two neighboring states.

Leader of the Islamic Revolution Ayatollah Seyyed Ali Khamenei talks with Russian President Vladimir Putin (c) in Tehran.

Figure 4. A screenshot of LocalSavvy

3. LOCALSAVVY

In this paper we propose a new paradigm for news aggregation according to the news sources associated with the stakeholders. The association between news sources and stakeholders is established through locations. Many of the entities in news articles are actually locations, such as cities and countries. The news articles published at the location provides the local views about the news issue. For other types of entities, such as people and organizations, we approximate their views with articles from the news sources that collocate with the entities. This approximation is based on the observation that local news sources usually speak in favor of the entities that are members of the local social group. By exploring the differences in news source locations, the new paradigm provides users with the capability to aggregate and browse various local points of view for the news issues in which they are interested. Furthermore, we can distinguish official news releases from unofficial news sources. The distinction can provide insight into the differences between the governments' perspectives and those of the general public.

We implement the proposed paradigm with a server-based prototype system called LocalSavvy. Users provide URLs of the news in which they are interested and LocalSavvy finds and aggregates news articles from the local news sources associated with the stakeholders of the news issues. Figure 4 shows a screenshot of the news results for "Putin's visit to Iran" presented by LocalSavvy.

When a news web page is given to the system, LocalSavvy identifies the entities mentioned in the news story as the stakeholders and finds the associated locations of theses stakeholders. For each location, the system formulates a set of queries to Google News Search [12] to gather news articles published in that location as well as official news releases from the local governments. LocalSavvy aggregates the results from Google News Search and organizes them according to the locations they belong to. Furthermore, LocalSavvy extracts the local opinions from the web pages and summarizes them in the news results. The summaries help users to browse the different points of view in the aggregated news articles. Moreover, when a user clicks on one of the search results, the local opinions are highlighted in the news web page, as shown in figure 4.

4. IMPLEMENTATION OF LOCALSAVVY

Figure 5 shows the system architecture of LocalSavvy. The server-based system accepts as input the URL of the news article in which the user is interested. The *News Story Modeler* processes the news web page and creates a model about the news issue. Based on the news model, the *News Aggregator* identifies the stakeholders and infers the related locations by consulting the *Knowledge Base*. The News Aggregator then formulates a set of queries to the search engine to find news articles published at the related locations. The news articles are aggregated according to the locations and presented to the user. The *Opinion Extractor* identifies and highlights the local opinions in the news web pages of the aggregated results. A separate *Knowledge Acquisition* module is responsible for creating and maintaining the knowledge base by searching the web. The following subsections describe each component in detail.

4.1 News Story Modeling

The main task of News Story Modeler is to extract the key information of the news article, which provides the bases for

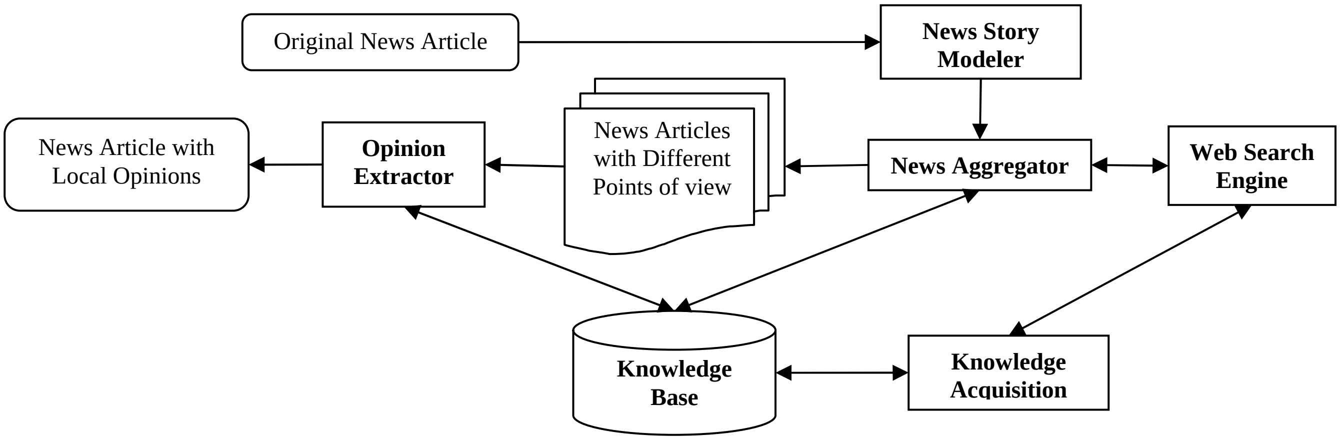


Figure 5. System architecture

query formulation. To this purpose, the system retrieves the web page with the URL specified by the user and extracts the news content of the web page using a method similar to that of Ma et al. [20]. According to previous research [16, 19, 30], named entities provide important information for news events and topics. Therefore, our system models news stories with two term vectors, one consisting of all the named entities and the other consisting of other non-entity words.

LocalSavvy tags named entities in news articles with the web service provided by ClearForest Semantic Web Services (SWS) [9]. A benefit for using SWS is that it provides some support for coreference resolution, which is critical for creating the vector of named entities. We supplement the coreference resolution of SWS with our own procedure. It works as follows, for two entities of the same type, if the tokens of one entity are totally contained within the other, or the name of one entity is the abbreviation of the other, the two entities are treated as coreference. In addition, nationality is treated as support for the country it belongs to.

To create the vector representations, we adopted the modified TF-IDF model for term weighting presented by Liu et al. [19]. The mechanism is based on the insight that important information is presented near the top in news writing [18, 26]. The modified Term Frequency (TF) is calculated from both the frequency and position of word occurrence. The weight of each word is the sum of the scores for all of its occurrences, with word occurrences in the title or near the top of the article being assigned higher scores. The named entities are weighted with the modified TF. Other non-entity words are weighted with modified TF-IDF, with IDF computed from an archive of 343,187 news stories.

4.2 Aggregating Different Points of view

The goal of LocalSavvy is to aggregate points of view from local news sources associated with the stakeholders. The first step for the News Aggregator is to identify the related locations of the news issue. From the vector of named entities produced by the news story modeler, the system collects the locations, people, and organizations as the stakeholders, with their importance ordered by their weights in the vector. For the specific locations mentioned in the original article, LocalSavvy further infers more generic locations by consulting its knowledge base. For example, the country “United States” will be added as a related location if the city “Washington” is mentioned in the original article. This generalization improves the system’s recall in case local articles from the specific locations are not available. For entities other than locations, LocalSavvy finds the related locations of the people or organizations in its knowledge base. For instance, state officials (e.g. presidents, foreign ministers, etc) are matched to the countries that they belong to, and companies are matched to the places in which they are located.

In order to find the relevant news articles from the related locations, LocalSavvy utilizes Google News Search. The search engine provides various advanced search options that allows the system to create very specific searches. Query formulation is a three-step process. First, the system constructs queries by combining the top named entities and non-entity words from the two vector representations produced by the News Story Modeler. Second, the system binds the queries with a time window around the publication dates of the original news article. Third, the system generates two sets of queries separately for official news releases and general public views. For the official news releases, the system consults its knowledge base for the domain names of the local government websites and restricts the search to those official websites. For the opinions of the general public, the system collects the local news articles by restricting the search to news sources in the locations.

We implemented a fallback strategy for news aggregation. If too few results are returned for a website or location, the system will gradually relax the query by removing search terms. Because named entities are important for identifying related news articles [16, 20], our system preferably preserves the named entities in the query and first deletes the snon-entity words with lower weights.

4.3 Location Related Knowledge Acquisition

In story modeling and searching, LocalSavvy requires knowledge about the relationship between non-location entities and locations. For example it needs to know which country the person came from or where is the organization located at. Instead of using a pre-built knowledgebase, LocalSavvy autonomously acquires the knowledge from the web. When an information request sent to the knowledge base is not satisfied, the system will automatically query the web search engine to find the answer and cache it for future use. The system also periodically updates its knowledge base with the web to correct or eliminate out-of-date information.

The problem arises of identifying the correct locations of people and organizations with potentially ambiguous names. We implements a mechanism similar to Question Answer (QA) systems which extracts knowledge from results of web search engine. Specifically, LocalSavvy utilizes Google News Search to acquire the knowledge it needs. Since a stakeholder will appear in multiple news reports of an event over the time span of that event, searching for the entity names in news search engine restricted within a certain time period can eliminate a lot of ambiguity. Furthermore, news articles are written in relatively standard format, resulting in effective information extraction.

Google News Search provides text snippets that contain the search terms in the web page as the summaries for search results. Instead of retrieving the full page, our system uses the summaries to extract the answers, enabling real-time knowledge acquisition. The summaries are tagged with a Brill Tagger [2] and the locations related to the entities are extracted from the summaries according to a set of linguistic rules. For example:

PERSON | ORGANIZATION (noun)* prep LOCATION

LOCATION | NATIONALITY (noun)* PERSON

Similar to QA systems [3, 17], LocalSavvy implements a voting mechanism to improve the precision of knowledge acquisition. The system clusters similar answers and selects the one with the most votes. Our system depends on the data redundancy on the web to acquire fast answers, which is shown to be very effective in our experiments (section 5.2).

4.4 Opinion Extraction

After LocalSavvy finds the local news articles, the system retrieves the web pages and identifies the local opinions. Our opinion extraction is distinctive in the way that it only extracts local opinions, i.e. opinions from the entities collocated with the news sources. This process involves two steps: first, identify the opinion sentences in the web page; second, extract the opinion holders. If the opinion holder is related to the location according to the knowledge base, the corresponding opinion is identified as a local opinion.

Previous studies indicated that subjective clues are especially effective for opinion identification [14, 24, 29]. To extract opinion sentences from news articles, we utilized FrameNet [1] data to derive the subjective clues. FrameNet is a lexical database of semantically hand-annotated data based on Frame Semantics [10]. A semantic frame is a conceptual structure that describes a particular type of situation. Each semantic frame consists of a set of lexical units and some frame elements. We went through the 884 frames in FrameNet and selected all the opinion-related frames. Then we collected all the frame evoking verbs in the opinion-related frames and use the verbs as our subjective clues. Altogether we collected 47 frames and 342 verbs. Table 1 shows examples of the opinion-related frames and the subjective clues.

Table 1. Examples of opinion-related frames

Frame	Subjective clues
Agree_or_refuse _to_act	agree, decline. refuse
Appeal	appeal, plead
Complaining	complain, gripe, grouse, grumble, lament, moan, whine
Predicting	claim, forecast, foretell, predict

The subjective clues are stemmed with the Porter stemmer [27] and searched in the article's sentences. The sentences containing the subjective clues are identified as opinion sentences. After opinion identification, the system selects the opinion holders from the persons and organizations within the opinion sentences with a set of extraction rules. The opinion holders are then queried in the knowledge base to find the associated location. If the location matches the source location of the news article, the sentence is highlighted as a local opinion sentence.

5. EXPERIMENTS

5.1 User Study

The goal of LocalSavvy is to aggregate local points of views about the news issues in which users are interested. Evaluating the efficacy of the system requires subjective judgment about whether the aggregated news articles express relevant local opinions. Thus, we conducted a user study to analyze the news results aggregated by LocalSavvy. Specifically, we investigated the following questions in our user study:

1. How relevant are the news articles to the news issue in which users are interested?
2. How different are the points of view presented by the news

```
<test-set>
        <topic> Uneasy allies in historic summit </topic>
        <background-story> http://news.bbc.co.uk/2/hi/europe/7045713.stm </background-story>
        <stakeholder>
                <entity> Iran </entity>
                <local-view> Leader Stresses Iran-Russia Ties </local-view>
                <local-view> Iran, Russia sign joint statement </local-view> </stakeholder>
        <stakeholder>
                <entity> United States </entity>
                <local-view> Vladimir Putin's Iran Trip in Historical Perspective</local-view>
                <local-view> Putin: No proof Iran is seeking nukes</local-view> </stakeholder>
        <stakeholder>
                <entity> Russia </entity>
                <local-view> Israeli PM Olmert to talk Iran, Mideast in Moscow Thursday </local-view>
                <local-view> Putin to support Iran Six policy on Tehran </local-view> </stakeholder>
</test set>
```

Figure 6. An exemplar test set

articles published at different locations?

3. How well do the news articles from local sources represent the opinions of the related stakeholders?

To prepare the test collection of news articles, we ran LocalSavvy on the "World News" RSS feed of Google News on October 15, 2007. Within the 20 news items in the RSS feed, for 9 news items LocalSavvy was able to find local news articles from more than three stakeholders. The 9 news items were presented to the subjects as the *background story*. For each news item, we gathered the top two search results from the top three locations and used them as the *test articles*. When there were official news releases as well as general news articles for a location, we took the first news article from both categories. Consequently, we gathered 5 or 6 test articles for each topic (for some locations, the system only returned one news article). For each news issue, the background story and the test articles found by LocalSavvy comprise a test set. In total we collected 9 test sets with 50 test articles. Figure 6 shows an example of the test sets. It should be noted that the test articles were presented to the subjects as plain text documents with only the titles and content extracted from the news web pages. Therefore, the subjects were not aware of the sources of the test articles in the user study. Their judgments were purely based on the content of the news article.

Our user study involved 22 subjects, all students of a university electrical engineering and computer science department. All subjects were native English speakers. The subjects were asked to select 2 news issues that they were most interested in. Because of this free choice, the subjects were not divided evenly among the test sets. Each test set was selected by 3 to 10 subjects. The subjects read the background story and the test articles in the selected test sets. They were then asked to rate each test article along different dimensions and complete a questionnaire to provide reasons for their judgments. The following subsections report our study for the above three questions.

5.1.1 Relevance of retrieved news articles

The first criterion to be met for the gathered news results is that they should be relevant to the news issue in which the user is interested. The relevance of the aggregated articles indicates the efficacy of query formulation of the system. In our user study, we asked the subjects to judge the relevance of the test articles to the situation described in the background story. The subjects rated each test article from 1 to 5, with 5 being most relevant and 1 being irrelevant. The *relevance score* of each test article is computed by averaging the ratings of the subjects. Figure 7 shows the distribution of relevance scores of the 50 test articles.

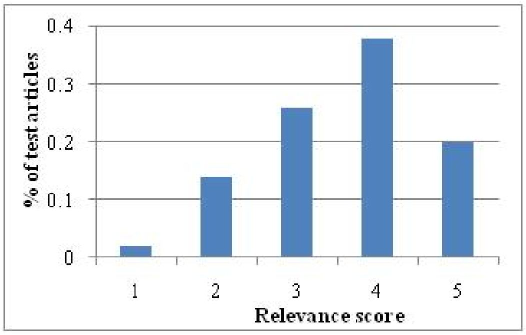


Figure 7. Distribution of relevance scores

As shown in figure 7, 84% of the test articles achieved a relevance score of at least 3. The mean relevance score of the 50 test articles is 3.98, with standard deviation 0.94. As discussed in section 4.2, LocalSavvy uses short queries bounded with publication dates to ensure high recall. Because of the fallback strategy, some of the queries consisted only of named entities. This test demonstrates that the short queries formulated by LocalSavvy are effective for finding relevant news articles.

5.1.2 Differences of local points of view

LocalSavvy explores the differences in news source locations to aggregate different points of view. To test whether the news articles published in different locations actually present different views, we asked the subjects to randomly select two pairs of test articles for each news issue they chose to read and rate the differences of the articles in terms of the points of view expressed. In total there are 88 pairs of articles: 23 pairs are from the same location (referred to as same location pairs) and 65 are from different locations (referred to as different location pairs).

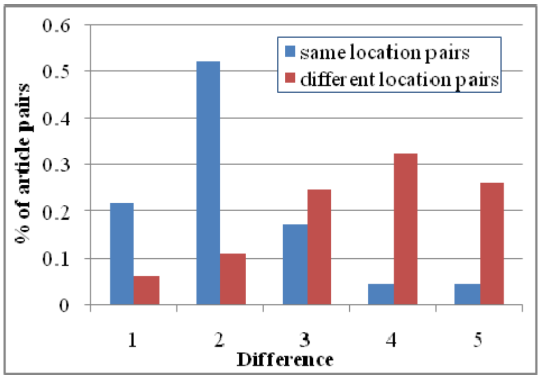


Figure 8. Distribution of difference ratings

The subjects rated the difference of the perspectives expressed in the article pairs from 1 to 5, with 5 being radically different and 1 being exactly the same. Figure 8 illustrates the distribution of the difference ratings for the two kinds of article pairs. The mean rating for the same location pairs is 2.17; and the mean rating for the different location pairs is 3.62, much higher than the same location pairs. As shown in figure 8, most of the articles published at the same location were judged to present similar points of view. In contrast, the subjects were able to perceive different perspectives in the articles published at different locations. The study demonstrates that the differences in news source locations are meaningful for exploring different points of view.

5.1.3 Degree of partiality of local news

The previous subsection demonstrates that news articles published at different locations are likely to yield different perspectives. Then how well do the news articles aggregated by LocalSavvy represent the local points of view? We explored this third question with a classification task in the user study. We presented the test articles to the subjects as plain text documents with only the titles and content of the news articles. The subjects were asked to classify each test articles into one of the three related locations of the background story, according to the points of view expressed in the test articles. If the test article represents the perspectives of the stakeholder strongly enough to be clearly perceived by the readers, the article should be consistently assigned to the correct stakeholder.

In our study, the subjects generated ranked lists of the stakeholders for the test articles, with highest rank of 2 and lowest rank of 0. For each test article, we computed its *degree of partiality* by averaging the ranks of the real stakeholder in the ranked lists generated by the subjects for that test article. Consequently, the test article that best represents the stakeholder's perspective will have the degree of partiality of 2, and the least will have the degree of partiality of 0.

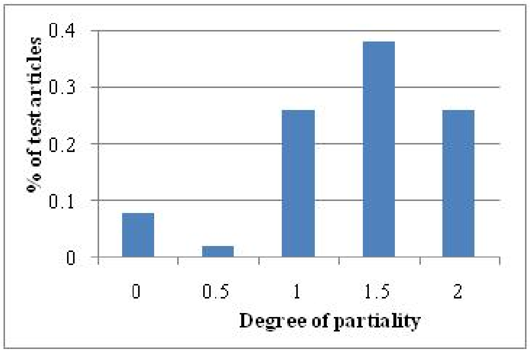


Figure 9. Distribution of degrees of partiality

Figure 9 illustrates the distribution of degrees of partiality for the 50 test articles. The mean degree of partiality for the 50 test articles is 1.51, with standard deviation of 0.52. As shown in figure 9, a large portion of the test articles express the opinions of the stakeholders clearly enough to be perceived by the readers. Our analysis finds that the test articles with very low degrees of partiality are mostly articles from the least important stakeholders, which have the lowest weights in the news models. For example, the two test articles with the lowest degrees of partiality are from entities mentioned only once near the end of the background stories as additional facts. They are not really "stakeholders" involved in or affected by the events. It was also pointed out by our subjects that "*(the test article) is less relevant to worldwide readers because it (the entity) is not a principal actor in this issue*". If we exclude the test articles from the least important stakeholder and only consider test articles from the top two stakeholders, the mean degree of partiality for the rest of 34 test articles is 1.69, with standard deviation of 0.33, which is higher than the overall mean. This finding supports our intuition that important entities of news issues are most likely to express strong opinions. It also indicates that LocalSavvy should be more selective about the stakeholders.

5.2 Knowledge Acquisition Evaluation

LocalSavvy relies on its knowledge base for news analysis and opinion extraction, whose knowledge is dynamically acquired from the web. The knowledge acquisition module utilizes the summaries returned by web search engines to find the associated location of people and organizations in real-time. We evaluated the performance of knowledge acquisition with two test sets of person-location pairs. The first test set was collected from Wikipedia's "List of current heads of state and government" [28]. We extracted the heads of state (e.g. presidents and kings) and heads of government (e.g. prime ministers) of each country in the list. We created the test pairs with the person names and their related countries. Altogether there are 302 persons from 179 countries in the first test set. To evaluate the accuracy in finding information about less famous persons, we invited an undergraduate student, who is not related to our project, to independently create the second test set from news articles. The student hand annotated 24 news articles from news websites and assembled a list of 53 persons, belonging to 33 different places. The second test sets includes a variety of types of person, such as spokesman, attorney, campaigner, etc.

The knowledge acquisition module queried the person name in each test pair with Google News Search and identified the related location of the person using the summaries of the first 10 results returned by the search engine. We evaluated the results with two measurements: coverage and accuracy. *Coverage* is the percentage of test pairs with search results within the whole test set; *accuracy* is the percentage of correctly identified test pairs among the pairs with search results. Table 2 shows the performance of knowledge acquisition on the two test sets.

Table 2. Performance of knowledge acquisition

	Coverage	**Accuracy**
Test set 1	92.7%	82.1%
Test set 2	96.2%	76.4%

As shown in table 2, the news search engine provides a large coverage both for the high profile state leaders and for the less famous people appearing in recent news. Moreover, there are a large number of news articles about any single news event. Therefore, this approach to knowledge acquisition is capable of yielding accurate results despite its simplicity.

6. FUTURE WORK

Our user study suggests several ways to improve local opinion retrieval. As shown in the user study, the entities mentioned only peripherally in news articles did not produce interesting opinions about the issue. Accordingly, we plan to refine stakeholder selection in order to focus local news retrieval on more important actors of the news issues. It was also reported by our subjects that factual news articles did not present distinct points of view. To filter out the articles that only consists of objective facts, we plan to build a classifier to distinguish the factual news reports from opinionated articles. Finally, we are exploring machine learning techniques for ranking the retrieved results according to their degree of partiality using features like major opinion holders and first person reference.

We implemented a primitive opinion extraction mechanism in the current prototype system. As part of our future research plan, we will extend this module to utilize more syntactic and semantic information to enhance the accuracy of opinion identification and opinion holder extraction.

7. CONCLUSION

In this paper, we propose a new paradigm for aggregating local news articles according to news sources locations. As studied in journalism and also demonstrated in our user study, news articles produced by different social groups present different attitudes towards and interpretations of the same news events. Finding and aggregating news articles from locations associated with stakeholders provides users insights into the local points of view. We implement this paradigm in LocalSavvy. LocalSavvy models the news articles provided by users. Based on the news story model, it finds the locations associated with the stakeholders and queries Google News Search for official news releases and general public news articles from those locations. LocalSavvy further extracts the local opinions in the retrieved results, which are presented in summaries and highlighted in the news web pages. The system acquires location related knowledge from the web. Our experiment shows that the knowledge acquisition module performs satisfactorily with the wealth of information on the web. We evaluate our system with a user study. The quantitative and qualitative analysis shows that the news articles aggregated by LocalSavvy present relevant and distinct opinions of the stakeholders, which can be clearly perceived by the subjects.

Our research explores the meaningful differences in information sources locations in the context of news reading. This new paradigm for news aggregation provides readers the capability to browse the various local points of view about the news issues, which are interesting and helpful for individual readers, organizations and governments.

8. REFERENCES

[1] Boas, H. C. "From Theory to Practice: Frame Semantics and the Design of FrameNet". In Semantisches Wissen im Lexikon. Tübingen: Narr. 2005

[2] Brill, E. “A simple rule-based part of speech tagger”. In Proce.of the 3rd conference on Applied natural language processing. 1992

[3] Brill, E., Dumais, S., Banko, M. “An analysis of the AskMSR question-answering system”, In Proc. of EMNLP. 2002

[4] Budzik, J. and Hammond, K. J. “User Interactions with Everyday Applications as Context for Just-in-time Information Access”. In Proc. of IUI. 2000

[5] Budzik, J., Hammond K., and Birnbaum, L. “Information access in context”. Knowledge based systems 14 (1-2). 2001

[6] Chen, Y., Di Fabbrizio, G., Gibbon, D., Jana, R., Jora, S., Renger, B., and Wei, B. "GeoTracker: Geospatial and temporal RSS navigation", in Proc. of the 16th WWW. 2007.

[7] Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. “Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns”. In Proc. of HLT/EMNLP-05. 2005.

[8] Christel, M. G., Olligschlaeger, A. M., and Huang, C., “Interactive Maps for a Digital Video Library”, IEEE Multimedia. 2000

[9] ClearForest Semantic Web Services (SWS) http://sws.clearforest.com/. 2007

[10] Fillmore, C. J. “Frame semantics and the nature of language”. In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, Volume 280: 20-32. 1976

[11] Finkelstein, L., Gabriolovic, E., Matias, Y., Rivilin, E., Solan, Z., Wolfman, G., and Ruppin, E. “Placing search in context: The concept revisited”. In Proc. of WWW. 2001

[12] Google News. http://news.google.com/. 2007

[13] Kim, S-M. and Hovy, E. “Identifying Opinion Holders for Question Answering in Opinion Texts”. In Proc. of AAAI Workshop on Question Answering in Restricted Domains. 2005

[14] Kim, S-M. and Hovy, E. “Extracting opinions, opinion holders, and topics expressed in online news media text”. In Proc. of ACL/COLING Workshop on Sentiment and Subjectivity in Text. 2006.

[15] Ku, L-W., Liang, Y-T., and Chen, H-H. “Opinion extraction, summarization and tracking in news and blog corpora”. In Proc. of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs. 2006

[16] Kumaran, G. and Allan, J. “Text classification and named entities for new event detection”. In Proc. of SIGIR. 2004

[17] Kwok, C. C. T., Etzioni, O., and Weld, D. S. “Scaling question answering to the Web”. In Proc. of WWW. 2001

[18] Lin C-Y and Hovy E. “Identify Topics by Position”. In Proc. of the 5th Conference on Applied Natural Language Processing. 1997

[19] Liu, J., Wagner, E., and Birnbaum L. “Compare&contrast: using the web to discover comparable cases for news stories”. In Proc. of 16th WWW. 2007

[20] Ma, L., Goharian, N. and Chowdhury, A. “Automatic Data Extraction From Template Generated Web Pages”. In Proc. of PDPTA 03. 2003

[21] Perry, D.K. “News reading, knowledge about, and attitudes toward foreign countries”. Journalism Quarterly, 67(2). 1990

[22] Radev, D. R., Qi, H., Zheng, Z., Blair-Goldensohn, S., Zhang, Z., Fan, W., and Prager, J. “Mining the web for answers to natural language questions”, In Proc. of the 10th CIKM. 2001

[23] Rhodes B. J. and Maes, P. “Just-in-time information retrieval agents”. IBM System Journal 39(4): 685-704. 2000

[24] Riloff, E. and Wiebe, J. “Learning extraction patterns for subjective expressions”. In Proc. of EMNLP. 2003.

[25] SmugMug Inc., “SmugMaps: combining the power of Google Maps with 65,000000+ SmugMug photos”. http://maps.smugmug.com/. 2006

[26] van Dijk, T. A. “News as discourse”. Hillsdale, NJ: Lawrence Erlbaum. 1988

[27] van Rijsbergen, C.J., Robertson, S.E. and Porter, M.F. “New models in probabilistic information retrieval”. British Library Research and Development Report, no. 5587. 1980

[28] Wikipedia. http://www.wikipedia.org/. 2007

[29] Wiebe, J., Wilson, T., and Bell, M. “Identifying Collocations for Recognizing Opinions”. In Proc. of ACL/EACL Workshop on Collocation. 2001

[30] Yang, Y., Zhang, J., Carbonell, J., and Jin, C. “Topic-conditioned novelty detection”. In Proc. of SIGKDD. 2002