NewsFinder: Automating an Artificial Intelligence News Service

Liang Dong[1], Reid G. Smith[2], Bruce G. Buchanan[3]

[1] School of Computing, Clemson University, SC, USA, ldong@clemson.edu
[2] Marathon Oil Corporation, rgsmith@marathonoil.com
[3]Computer Science Department, University of Pittsburgh, PA, USA, buchanan@cs.pitt.edu

Abstract

NewsFinder automates the steps involved in finding, selecting and publishing news stories that meet subjective judgments of relevance and interest to the Artificial Intelligence community. NewsFinder combines a broad search with AI-specific filters and incorporates a learning program whose judgment of interestingness of stories can be trained by feedback from readers. Since August, 2010, the program has been used to operate the *AI in the News* service that is part of the AAAI AITopics site.

Task Description

Selecting a small number of interesting news stories about AI, or any other topic, requires more than searching for individual terms. Since it is time-consuming to find and post interesting stories manually, we have designed and written an AI program called NewsFinder that automatically collects news stories from selected sources, rates them with respect to a learned measure of goodness, and publishes them as the *AI in the News* service that is part of the AAAI AITopics[1] site described in (Buchanan, Glick and Smith 2008).

The goal for NewsFinder is to publish a small, select set of articles that are of general interest to the AI community. The task is similar to asking Google News to find stories about AI, but differs in several respects. Google News is driven by users' queries to find stories containing a set of keywords from thousands of news sources. By contrast, NewsFinder uses multiple pre-determined queries and RSS feeds to find stories about AI from a select set of highly credible sources. Both systems use a variety of means to rank stories. For Google News, the signals include the number of user clicks, the estimated authority of a publication in a particular topic, freshness, geography, etc. NewsFinder uses publication authority and freshness. It also estimates how much the content discusses issues of interest to the AI community. Both systems cluster similar stories (by different methods) and both incorporate feedback from users (analogous to Netflix or Amazon). NewsFinder uses these signals to learn and continually retrain its ranking component. An empirical comparison of the two systems is given below.

For each news story, NewsFinder determines which particular aspect of AI is the main focus of the story (e.g., robots, machine learning, and applications). It also detects and removes near-duplicate stories that cover the same event.

Design

The work of the NewsFinder application consists of two major tasks: Daily Crawling and Weekly Publishing, as shown in Figure 1 and described below.

News Story Representation

As is common in information retrieval applications, we use Salton's weighted term-frequency vector space model (Salton and Buckley 1988) to represent each news story. Let $W = \{f_1, f_2, ..., f_m\}$ be the complete vocabulary set of the crawled news after stemming, morphing and stoplist filtering. On average, each story contains about 200 different terms. The term frequency vector X_i of news story d_i is defined as

$$X_i = [x_{1i}, x_{2i}, ..., x_{mi}]^T$$

$$x_{ji} = \log(t_{ji} + 1) \cdot \log(\frac{n}{idf_j})$$

where t_{ji} denotes the frequency of the term $f_j \in W$ in the news story, d_i; idf_j denotes the inverse of the number of stories containing word f_j; and n denotes the total number of candidate stories in the database. In addition, X_i is normalized to unit Euclidean length.

[1]http://www.aaai.org/AITopics/pmwiki/pmwiki.php/AITopics/AINews

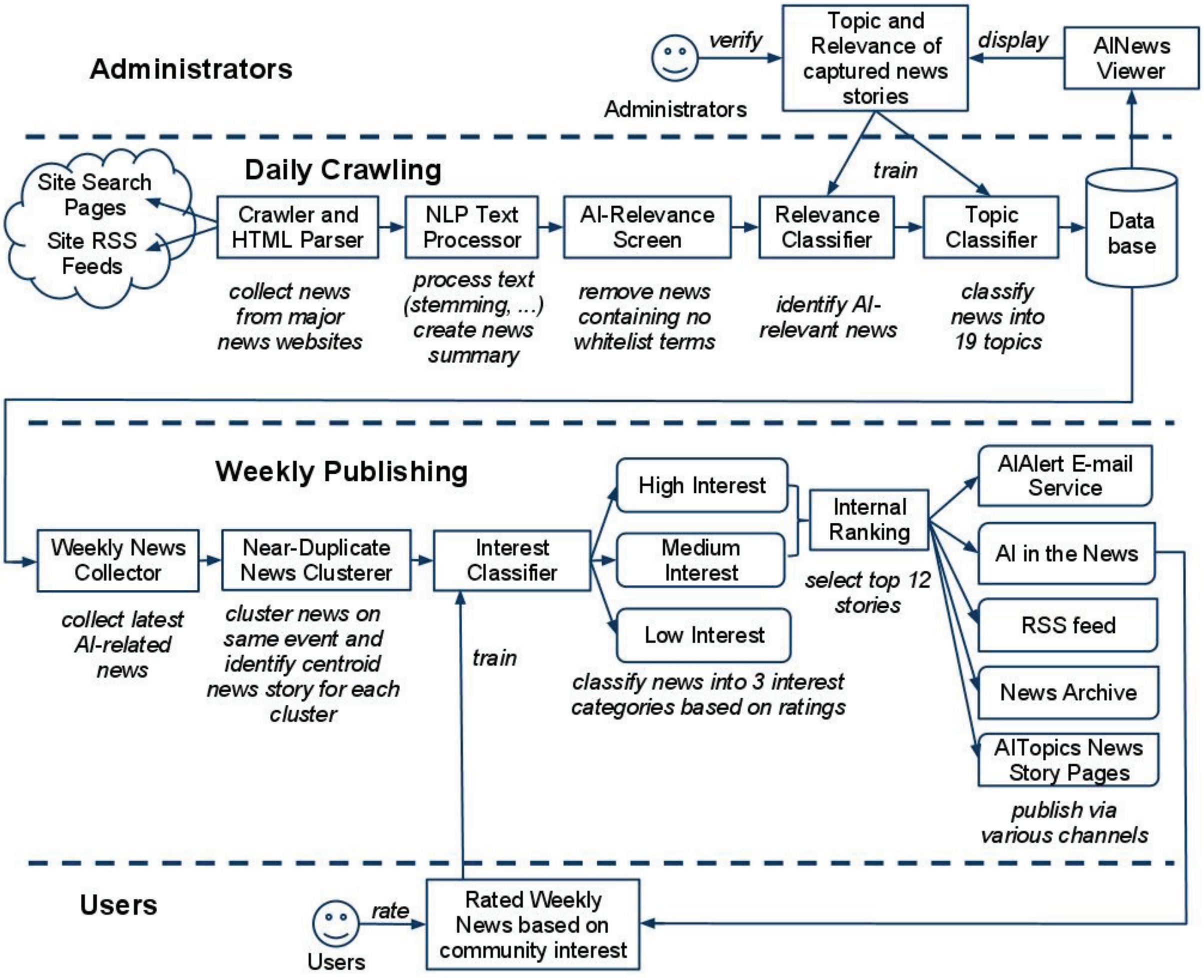

Administrators
verify
Topic and Relevance of captured news stories
display
AINews Viewer
Administrators
Daily Crawling
train
Site Search Pages
Site RSS Feeds
Crawler and HTML Parser
NLP Text Processor
AI-Relevance Screen
Relevance Classifier
Topic Classifier
Data base
collect news from major news websites
process text (stemming, ...) create news summary
remove news containing no whitelist terms
identify AI-relevant news
classify news into 19 topics
Weekly Publishing
High Interest
Medium Interest
Low Interest
Internal Ranking
select top 12 stories
AIAlert E-mail Service
AI in the News
RSS feed
News Archive
AITopics News Story Pages
publish via various channels
Weekly News Collector
Near-Duplicate News Clusterer
Interest Classifier
collect latest AI-related news
cluster news on same event and identify centroid news story for each cluster
train
classify news into 3 interest categories based on ratings
Users
rate
Rated Weekly News based on community interest
Users

Figure 1 The NewsFinder Program. The two main tasks are (a) Daily Crawling for candidate stories and (b) Weekly Publishing of top-ranked stories. Learning from both administrators and users is a central aspect of its operation.

Daily Crawling

News Sources

Since we value quality over quantity, NewsFinder is restricted to 29 news sources, chosen for their national or international scope, credibility and stability. These sources are divided into three groups: (1) *general sources*, including BBC, CNN, Discovery, Forbes, The Guardian, LA Times, MSNBC, Nature, New York Times, NPR, USA Today, Wall Street Journal, Washington Post; (2) *technology sources*, including CNet, MIT News, New Scientist, Popular Science, Scientific American, Wired, and ZDNet, which focus on science and technology; and (3) *AI sources*, such as Kurzweil.Net[2] and Robots.Net[3], which focus on news about artificial intelligence.

HTML Parsing and Content Extraction

If a source has an in-site search function, NewsFinder queries it to find stories that contain the keywords 'artificial intelligence', 'robots', or 'machine learning'. If a source provides RSS feeds, then NewsFinder selects available feeds labeled 'AI', 'robots', 'technology' or 'science'. Stories verified as 'blog posts' or 'press releases' are skipped. One or two thousand stories are collected each week from these sources.

To retrieve the content of news pages, we designed a generic HTML parser and a set of derived parsers customized for each news source. This method improves the quality of the extracted news text by eliminating user comments, navigation bar/menu text, JavaScript, videos, and irrelevant in-site hyperlinks from the original HTML. Additional customized parsers can be added into the pool

[2] http://www.kurzweilai.net/

[3] http://robots.net/

with little difficulty, but with a need for maintaining them as sources change their formats.[4]

A parser extracts four items associated with each news story: URL, title, publication date and story text. A news story is skipped if the publication date is outside the crawling period (currently seven days). Although some stories spread over multiple web pages, we only crawl the first page of text on the assumption that it adequately reflects the content of the whole story.

Text Processing and Summarization

NewsFinder next performs several natural language processing steps on each story using the Natural Language Toolkit (Loper and Bird 2002). These include collecting n-grams (n=1, 2, 3) for *whitelist* filtering, lower-capitalizing, stemming, and stoplist removal. We use a stoplist of 583 common words, modified from the list at MIT[5]. We check for terms of particular interest, like 'Turing', by adding them to the *whitelist* described in the next section.

We cannot directly use descriptions from RSS feeds or in-site search results to generate concise summaries for stories because most of those descriptions contain incomplete sentences. Instead, we modified an existing Python summarizer[6] which extracts four sentences from the story text to describe the highlights that make the story interesting. The goal of the algorithm is to measure *tf-idf* over the entire story text, and then select the four sentences that contain the most frequent terms. In the end, it re-assembles the selected sentences in their original order for readability.

AI-Relevance Screening

Because we use RSS feeds under general keyword terms like 'technology' or 'science', many returned stories are not relevant to artificial intelligence at all.

A two-level screening strategy is used to filter irrelevant news. At the first level NewsFinder references a *whitelist* consisting of 64 n-grams whose inclusion in a story is necessary for further consideration. In addition to the terms 'artificial intelligence' and 'AI', the *whitelist* includes several dozen other unigrams, bigrams and trigrams that indicate a story has potential interest beyond the search terms. For example, mention of 'autonomous robots' makes a story more likely to interest the AI community than mere mention of 'robots' (which may be tele-operated). The stories passing the first level screening are saved into the candidate stories database.

Relevance: Binary Classifier

The second level screening is performed by a Support Vector Machine (SVM) classifier that is trained to give a binary judgment about relevance.

Support Vector Machines (Burges 1998) are widely used for supervised learning for classification, regression or other tasks by constructing a hyperplane or set of hyperplanes in a high dimensional space. They have been applied with success in information retrieval problems particular to text classification. We have used the LibSVM (Chang and Lin 2001) open source library to implement the Relevance classifier as well as the Ranking classifier (described in a later section).

The administrators (AI subject matter experts) provide feedback as training data for the Relevance classifier using the AINews Viewer tool[7] which enables an administrator to view the metadata and the source web page for a story, and to verify or change the assigned topic and/or relevance.

Topic Recognition

NewsFinder determines the main topic of each story among the 19 top-level topics[8] in AITopics. We implemented the topic classifier as a centroid-based classifier similar to that described in (Han and George 2000). We manually scrutinized 284 introductory documents from all categories from both AITopics and Wikipedia. We then used the vector space model to compute and normalize the centroid of each category.

NewsFinder computes the cosine vector similarity between a news text and the nineteen centroids. These similarity measures enable the program to assign the most similar centroid as the story's topic by measuring the dot product of normalized vectors (Manning, Raghavan et al. 2008). As the theory and practice of AI changes, new terms such as 'Kinect' and 'semantic web' will continue to emerge and the centroid classifier will need to be retrained to categorize stories correctly.

Finally, NewsFinder saves the candidate news stories and their metadata into the candidate stories database for subsequent processing.

Weekly Publishing

Currently, about forty candidates pass the relevance screening each week. The next steps are to eliminate stories that duplicate news on the same subject and to rank and publish the top dozen stories.

[4] We have also implemented a general parser for parsing any news site, similar to previous work (Gupta, Kaiser et al. 2003; Song, Liu et al. 2004), which is used to parse stories from Google News. However, this implementation is still experimental.

[5] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[6] http://tristanhavelick.com/summarize.zip

[7] http://www.aaai.org/AITopics/html/ainews-viewer/news.php

[8] The current major topics are: AI Overview, Agents, Applications & Expert Systems, Cognitive Science, Education, Ethical & Social Implications, Games & Puzzles, History, Interfaces, Machine Learning, Natural Language, Philosophy, Reasoning, Representation, Robots, Science Fiction, Speech, Systems & Languages, Vision.

Detecting Near-Duplicate Stories

We consider three cases of near-duplicate news: (a) exact duplicates (e.g., the same story from one wire service); (b) similar rewrites (with identical or slightly modified sentences in common); (c) reports of the same event (e.g., separately written descriptions of an event).

NewsFinder measures the vector space cosine similarity of *tf-idf*, with a temporal coefficient μ that reduces the cosine term in inverse proportion to the number of days separating the stories.

$$\mu \cdot sim(news_1, news_2)$$

where the similarity of a pair of stories is reduced by a linear temporal coefficient μ whose value is between zero and one proportional to the number of days difference in the publication dates of the stories (min 1-3 days = 1, max 140 days = 0).

If the computed similarity value is greater than a cutoff (0.23, selected empirically), we regard the pair of stories as near duplicates. We assume that the similarity of news is transitive so that we can cluster near-duplicates. For each cluster, we compute the centroid and select the story closest to it. After this stage, almost no duplicates remain.

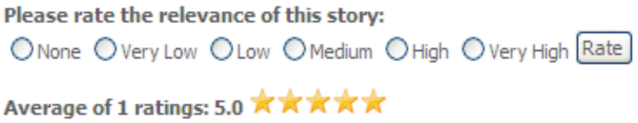
Two alternatives for detecting near-duplicates are the shingling algorithm described in (Broder, Glassman et al. 1997) and locality sensitive hashing in (Charikar 2002). (See (Kumar and Govindarajulu 2009) for recent progress in this field.) We chose to use the vector space model for its simplicity and speed, and its ability to detect near duplicate stories in case (c).

Ranking: Three-Tier Interest Classifier

We initially planned to use a Back Propagation Neural Network (BPNN) as a classifier to train the news, but we finally selected SVM. One reason is that recent studies (Dumais 1998; Joachims 1998; Zhang 2007; Zaghloul 2009) have shown that SVM leads to better text classification than BPNN, Bayes, Rocchio, C4.5 and kNN. (Joachims 1998) suggests four reasons for the superiority of SVM: (1) over-fitting protection in high dimensional input spaces; (2) few irrelevant features; (3) sparse document vectors; and, (4) most text categorization problems are linearly separable.

Another reason for choosing SVM is speed of implementation, due to the availability of LibSVM, a widely used, well-documented and efficient open source package. It is written in C++ with a Python-wrapped interface which is convenient for integration into NewsFinder.

We have built a multiclass classifier based on three "one against the rest" SVM-based probability classifiers to identify three interest categories as learned from the average of users' ratings. Stories with ratings 3.34-5.0 are classified "high" interest, those rated 1.67-3.34 are "medium" interest, and those rated 0-1.67 are "low" interest.



Please rate the relevance of this story:

○None ○Very Low ○Low ○Medium ○High ○Very High Rate

Average of 1 ratings: 5.0 ★★★★★

Figure 2. Rating Interface. Feedback from readers is used to retrain the SVM that classifies degree of interest to other readers.

A story is assigned to the category associated with the classifier that computes the highest probability. NewsFinder then discards stories classified into the low interest category.

The SVM classifiers are retrained every week, before new stories are published. The average rating for a story is used for retraining if there is general consensus among several raters, based on the standard deviation. Alternatively, an administrator's rating is used.

Stories in the same interest level category are then ranked by the following three factors:

- **Popularity:** A story published in more news sources is considered to have more general interest than a story appearing in just one source.
- **Whitelist Terms:** The more occurrences of *whitelist* terms in a story the more interesting it is considered to be.
- **Source:** A general source is considered to have more interest than a technology source, and that to have more interest than a specific AI source.

Once news stories in the high and medium interest categories have been ranked, NewsFinder selects the top 12 stories to publish, starting with stories in the high-interest category.

Retraining the Interest Classifiers

Training data are collected through the rating system, which is modeled after the five-star rating system used by Netflix (Herlocker, Konstan et al. 2004). Unlike Netflix, our purpose is to classify unseen items with respect to their likely interest to all readers of *AI in the News*, and not just their interest to the individual doing the rating. We assume that the aggregate of many ratings reflects the opinion of the community at large.

In Figure 2, when we collect feedback, we also show the average rating of other readers during the week, both as a number and as a row of stars, for readers who may wish to focus first on stories that others have rated highly.

The PmWiki Cookbook StarRater[9] is used to collect users' ratings. We record each user's rating and IP address for every news story. The IP address is a proxy for a user ID and allows us to record just one vote per news item per IP address. (We do capture user ID directly if the reader is logged into AITopics.) As with Netflix, if there are multi-

[9] PmWiki star rater http://www.pmwiki.org/wiki/Cookbook/StarRater

ple ratings for the same story from the same reader, only the last rating is used.

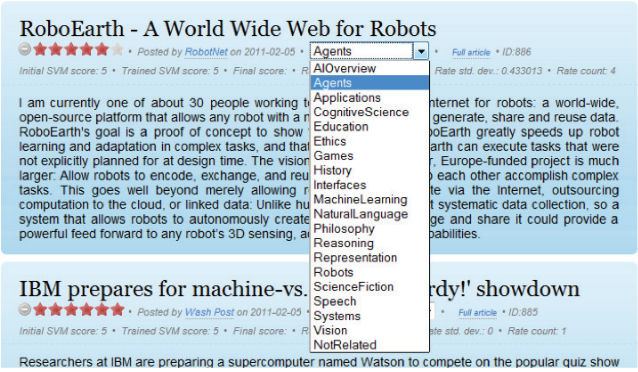
Publishing

The selected stories are published via five channels: (1) the Latest *AI in the News* summary page of AITopics, (2) the weekly "AIAlert" e-mail message to subscribers[10], (3) the RSS feeds associated with the AITopics major topics, (4) the AITopics news archive, and (5) individual story pages on AITopics.

Administration

Stories can also be added manually to be included in the current set of candidate stories in the Publication phase. Thus when an interesting story is published in a source other than the ones we crawl automatically, or is erroneously ranked below threshold for publication, it can still be published. It will also be included in subsequent training, which may help offset the inertia of training over the accumulation of all past stories collected under the program's own criteria.

Dozens of candidate stories are added to the database each week. Those stories need to be monitored and evaluated by administrators to continuously correct classification errors. Figure 3 illustrates the AINews Viewer tool that administrators use to view crawled news stories and their scores. The tool is deployed under the aaai.org domain outside of the PmWiki framework so that AJAX can be used to improve real-time interaction. Administrators can modify the Topic Classifier's result as well as rate the story. If an administrator sets the story's topic as 'NotRelated', that story will not be processed in the weekly ranking phase.



RoboEarth - A World Wide Web for Robots

• Posted by RobotNet on 2011-02-05 • Agents • Full article • ID:886

Initial SVM score: 5 • Trained SVM score: 5 • Final score: • R Rate std. dev.: 0.433013 • Rate count: 4

I am currently one of about 30 people working t nternet for robots: a world-wide, open-source platform that allows any robot with a n generate, share and reuse data. RoboEarth's goal is a proof of concept to show oEarth greatly speeds up robot learning and adaptation in complex tasks, and that arth can execute tasks that were not explicitly planned for at design time. The vision , Europe-funded project is much larger: Allow robots to encode, exchange, and reu each other accomplish complex tasks. This goes well beyond merely allowing r te via the Internet, outsourcing computation to the cloud, or linked data: Unlike hu t systematic data collection, so a system that allows robots to autonomously create ge and share it could provide a powerful feed forward to any robot's 3D sensing, a abilities.

IBM prepares for machine-vs. rdy!' showdown

• Posted by Wash Post on 2011-02-05 • Full article • ID:885

Initial SVM score: 5 • Trained SVM score: 5 • Final score: • R ate std. dev.: 0 • Rate count: 1

Researchers at IBM are preparing a supercomputer named Watson to compete on the popular quiz show

Figure 3. AINews Viewer

Next steps to improve the performance of the system include the following:

1. Encourage more users to contribute ratings;
2. Continue to use AINews Viewer to closely monitor the crawled news stories and their AI-relevance scores from the Relevance Classifier. Since the relevance classifier operates in a semi-supervised fashion, early misclassified result correction can greatly reduce mistakes for similar future stories;
3. Use the stories contributed manually by users to identify high-value additional news sources and new *whitelist* terms.

Validation, Use and Payoff

Preliminary validation studies (Dong, Smith and Buchanan 2011) before full deployment showed that the NewsFinder program was efficient and effective, and that retraining with users' feedback improved its performance. On a small sample, an administrator agreed with 84% of the program's selections and 100% of its omissions. We have experimented with many parts of the program and revised them somewhat as we have gained experience.

NewsFinder was put into routine operation in August, 2010 and the alerts service was made operational in December. Each week thousands of stories are read for relevance, about forty are ranked with respect to their interest, and a dozen or fewer are selected for publication. The *AI in the News* page is viewed by over 1000 readers each month (April, 2011). In addition, the AIAlert e-mail is sent every Monday morning to about 500 subscribers. Each week, story pages are added automatically to AITopics, where they remain searchable.

In one set of 351 stories that were manually examined, the procedure for detecting near-duplicates found two or more duplicates among 31 events (80 stories total) as shown in Table 1.

Table 1. Accuracy of duplicate detection, as judged by one administrator.

True Pos	False Pos	Precision	Recall	F1[11]
67	6	91.8%	77.9%	84.3%

Statistics on the use of the production system are shown in Table 2 below. Of all stories passing the relevance filters, 15%-20% have duplicates and all appear to have been successfully eliminated.

Among 249 stories that passed the relevance screening and duplicate elimination, and thus are scored with respect to interest, the overall rate of agreement between the program and an administrator is 61.8% on decisions to publish or not (threshold ≥ 3.0), with Precision = 0.813, Recall = 0.448 and F1 = 6.92. While the program recommended publishing about 70% of the stories passing the relevance

[10] To prevent spam generation, the AAAI distribution-list server does not allow automatic sending of e-mail. Thus a member of the AAAI staff must manually send a preformatted message to subscribers each week.

[11] Precision = TruePos/(TruePos+FalsePos)
Recall = TruePos/(TruePos+FalseNeg)
F1 = 2 (Precision Recall)/(Precision+Recall)

filters, the administrator recommended publishing about 60%. Agreement on the more highly rated stories (threshold ≥ 4.0), which more accurately reflects the subset to include in the best dozen stories, was 70% on these 249 candidates. In either case, most of the disagreement lies in NewsFinder's lower scores for stories that the Administrator would publish – which is acceptable given that we only want the best dozen to be posted anyway.

Table 2. Numbers (%) of stories rated over and under the publication threshold (≥ 3.0), by an Administrator and NewsFinder (N =249).

	Admin: Publish	Admin: Don't Publish
NewsFinder: Publish	65 (26%)	15 (6%)
NewsFinder: Don't Publish	80 (32%)	89 (36%)

The cost of developing the program was one student stipend for one and a half summers, plus volunteered time by two senior AI scientists. The PmWiki and Python systems are free and several free library packages have been incorporated into NewsFinder. Maintenance will be overseen by volunteers but we will need to hire programmers to consult on major problems when and if they arise. The monetary benefit is the saving of about ten hours a week of a webmaster's time, offsetting the 1.5 student stipends in a half year or less. Additional benefits accrue from the consistency and reliability of an automated service plus the unquantifiable benefits of providing useful information to the AI community.

Comparison with Google News

Probably the most widely used news retrieval service is Google News, which finds articles of interest published within the last 30 days. These are selected by keyword search and ranked with a proprietary algorithm. Because of its widespread use, we compare NewsFinder with this service.

However, considerable research is ongoing in other groups on learning to improve the effectiveness and efficiency of search engines, some of which has been published; e.g., (Yue and Joachims 2008), (Wagstaff et al. 2007), (Burges et al. 2006), (Tsochantaridis et.al. 2004).

Our problem closely resembles the problem described in (Yue and Joachims 2008) in which diverse topics may be included in the target set, although they address a larger diversity of topics than our limited set of twenty. As we believe also, Joachims notes that the key to success in searching specialized collections is to have a program that improves over time with use.

Table 3a. Top 12 news items listed in Google News by keyword 'artificial intelligence' at 10:00 p.m., Mar. 29, 2011 (GMT -5).

1	The Global Robotics Brain Project	IEEE Spectrum
2	Call for Papers: Expanding Human Boundaries: Cognitive Enhancement, AI and …	Inst. Ethics & Emerging Tech.
3	Woodlawn High Team Headed to National Robotic Competition	Patch.com
4	Transcending the flesh: the coming Singularity	GMANews.TV
5	Apple's iOS 5 to feature 'artificial intelligence' and voice control	Know Your Mobile(blog)
6	Global sales of robots to reach new heights in 2011	Vision Systems Design
7	Cougars tackle robotics challenge for national title	Ultimate Katy
8	I Took the Turing Test	New York Times
9	AI Artificial Intelligence (Blu-ray)	DVDTOWN.com
10	Japan brings artificial intelligence to rockets	PhysOrg.com
11	UC Davis to commercialize modular, mobile robots	ZDNet
12	Robots Are the Next Revolution, So Why Isn't Anyone Acting Like It?	IEEE Spectrum

Though the actual list of Google News sources is not known outside of Google, the stated information from Google is that it watches more than 4,500 English language publishers[12] and totally more than 25,000 publishers in 19 languages around the world.[13] The news sources are manually selected by human editors and Google has designed a customized news ranking algorithm that includes user clicks (popularity), the estimated authority of a publication in a particular topic (credibility), freshness and geography to determine which stories to show from the online news sources it watches.[14]

Google News can be characterized as a comprehensive and real-time news provider. It allows users to perform keyword searches on the latest news and build such searches into RSS feeds.

To illustrate the difference between Google News and NewsFinder, we conducted two keyword searches on 'artificial intelligence' and 'robot' (at 10 p.m., Mar. 29, 2011). The top 12 news items from each search are listed in Tables 3a and 3b respectively, comparing to 12 news items generated by NewsFinder in Table 3c. The time and date of the Google news search are arbitrary, and the result lists are not tuned.

As illustrated in Tables 3a and 3b, Google News collects most of the latest news, where 18 out of the 24 news articles were published in the most recent 24 hours. However, due to the direct keyword matching and variety of sources, a third of the news stories are not very interesting (e.g.,

[12] http://en.wikipedia.org/wiki/Google_News

[13] http://googlenewsblog.blogspot.com/2009/12/same-protocol-more-options-for-news.html

[14] http://searchengineland.com/google-news-ranking-stories-30424

Table 3b. Top 12 news items listed in Google News by keyword 'robot' at 10:00 p.m., Mar. 29, 2011 (GMT -5).

1	US sending robots to Japan to help nuclear plant	The Associated Press
2	Gizmo games for agile robots	The West Australian
3	Robot arm scoops goop stains like magic	CNET (blog)
4	Rockin' Robots	Polson Lake County Leader
5	The Virtual Doctor Is In: Robots In Hospitals	Investor's Business Daily
6	Global sales of robots to reach new heights in 2011	Vision Systems Design
7	Goshen middle schoolers gearing up for robot competition	WSBT-TV
8	Scientists show off a robot that can rebuild itself	Digitaltrends.com
9	This Isn't A Robot In A Suit With An iPad Head	Gizmodo Australia
10	Robot Lifeguard Emily Invented To Save Swimmers	Huffington Post
11	Rob Lowe: 'Tom Cruise is Like a Robot'	Showbiz Spy
12	Tuesday Robot: Jackie Chan, Chelsea Kane, Britney Spears	Robot Celeb

Table 3c.. Top 12 news items generated by NewsFinder, at 6:00 a.m., Mar. 30, 2011 (GMT -5).

1	Private EyeBot Lurks in the Shadows, Can Tail Suspects Without Being Seen	Popular Science
2	Spiders and crabs inspire robot locomotion	BBC News
3	UC Davis to commercialize modular, mobile robots	ZDNet
4	iRobot to the rescue	MIT News
5	Video: Robotic Swiss Quadrocopters Hold Their Own At Tennis	Popular Science
6	Let's Hope the Robots Are Nice	Wall Street Journal
7	New film on Alan Turing	Kurzweil
8	Friday Poll: Would you go under the robotic knife?	CNet
9	Delivering Results	Wall Street Journal
10	Tiny 3-D-Printed Insect Robots Take Flight	Wired
11	Move over, Einstein: machines will take it from here	Kurzweil
12	TEROOS robotic avatar gives your long-distance girlfriend a tiny, googly-eyed face (video)	Engadget

conference call for papers, review of the blue-ray version of the movie "A.I."). There are four stories about four high school robotic competitions (AI #3, #7, Robot #4, #7), and two stories are from gossip sites discussing movie stars (Robot #11, #12). In addition, AI #5 and Robot #9 might be of interest to certain readers, but not necessarily the overall AI Community.

"I Took the Turing Test" (AI #8) is a good story from a credible source. However, the story was published on Mar. 18, 2011, and was included in NewsFinder's weekly published list on Mar. 21 under the name "The Most Human Human".

There are two overlapping news articles from Google News and NewsFinder. One is a direct mapping from AI #11 to NewsFinder #3. Another is an indirect mapping from Robot #2 to NewsFinder #5 that describes the same event and refers to the same video, but is written by different authors. We prefer NewsFinder's choice since Popular Science is more international than The West Australian.

By comparing the news title, publisher and its content, we conclude that the current Google News can inform subscribers of the latest and comprehensive news, but it only maps keywords and doesn't filter irrelevant news by contents. By contrast, NewsFinder provides screened and refined weekly information to the AI community.

Conclusions & Future Work

Replacing a time-consuming manual operation with an AI program is an obvious thing for the AAAI to do, although intelligent selection of news stories from the web is not as simple to implement on a small budget as it is to imagine. There are many different operations, each requiring several parameters to implement the heuristics of deciding which dozen stories are good enough to present to readers.

NewsFinder has proved to be capable of providing a valuable service with low development cost. Off-the-shelf libraries make it possible to build an intelligent system in a short time, and reduce the problems of maintenance. Learning how to select stories that the community rates highly adds generality as well as flexibility to change its criteria as the interests of the community and the field itself change over time.

Besides continuously improving the crawling and ranking performance, we are interested in extending NewsFinder as a personalized weekly news service more like the Amazon and Netflix recommendation systems so that each person's news list is ranked individually based on his/her previous rating. In addition, by mining the news of interest to an individual reader, we can also discover and recommend the latest published journal articles by crawling the CiteSeer, IEEE and ACM websites.

Acknowledgments. We are grateful to Jim Bennett for many useful comments, especially on the Netflix rating system after which the NewsFinder rating is modeled, and to Tom Charytoniuk for implementing the initial prototype of this system.

References

Broder, A., Glassman, S., Manasse, M. and Zweig, G. 1997. Syntactic Clustering of the Web. *6th International World Wide Web Conference (WWW97)*.

Buchanan, B. G., Glick, J. and Smith, R. G. 2008. "The AAAI Video Archive." *AI Magazine*, **29**(1): 91-94.

Burges, J. C. C. 1998. "A tutorial on support vector machines for pattern recognition." *Data Mining and Knowledge Discovery* **2**(2): 121-167.

Burges, J. C. C., Ragno, R., and Le, Q. 2006. "Learning to rank with non-smooth cost functions." In proceedings of *the International Conference on Advances in Neural Information Processing Systems (NIPS)*.

Chang, C.-C. and Lin, C.-J. 2001. "LIBSVM: a library for support vector machines."

Charikar, M. S. 2002. "Similarity Estimation Techniques from Rounding Algorithms." In proceedings of *the 34th Annual ACM Symposium on Theory of Computing*.

Dong, L., Smith, R. G. and Buchanan, B. G. 2011. "Automating the Selection of Stories for AI in the News." In proceedings of *the 24th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems,* (*IEA-AIE '11)*, Syracuse, NY.

Dumanis, S., Platt, J., Heckerman, D., Sahami, M. 1998. "Inductive Learning Algorithms and Representations for Text Categorization." In proceedings of *the 7th International Conference on Information and knowledge management (CIKM '98),* Washington D.C., U.S.

Gupta, S., Kaiser, G., Neistadt, D. and Grimm, P. 2003. "DOM-based Content Extraction of HTML Documents." In proceedings of *the 12th International World Wide Web Conference (WWW2003)*, Budapest, Hungary.

Han, E.-H. S. and George, K. 2000. "Centroid-based document classification: Analysis and experimental results." In proceedings of *the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*.

Herlocker, J., Konstan, J., Terveen, L. and Riedl, J. 2004. "Evaluating Collaborative Filtering Recommender Systems." *ACM Transactions on Information Systems* **22**.

Joachims, T.. 1998. "Text categorization with support vector machines: Learning with many relevant features." In proceedings of *the 10th European Conference on Machine Learning, (ECML'98)*, Chemnitz, Germany.

Kumar, J. P. and Govindarajulu, P. 2009. "Duplicate and Near Duplicate Document Detection: A Review." *European Journal of Scientific Research* **32**(4): 514-527.

Loper, E. and Bird, S. 2002. "NLTK: The Natural Language Toolkit." In Proceedings of *the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Philadelphia, Association for Computational Linguistics.

Manning, D. C., Raghavan, P. and Schutze, H. 2008. "Introduction to Information Retrieval" Cambridge University Press.

Salton, G. and Buckley, C. 1988. "Term-weighting approaches in automatic text retrieval." *Information Processing & Management* **24**(5): 513-523.

Song, R., Liu, H., Wen, J.-R. and Ma, W.-Y. 2004. "Learning Block Importance Models for Web Pages." In proceeding of *the 13th International World Wide Web Conference (WWW 2004)*, New York.

Tsochantaridis I., Hofmann T., Joachims T., and Altun Y. 2004. "Support Vector Machine Learning for Interdependent and Structured Output Spaces." In proceedings of *the International Conference on Machine Learning (ICML)*.

Wagstaff, K., desJardins, M., Eaton, E., and Montminy, J. 2007. "Learning and visualizing user preferences over sets." *American Association for Artificial Intelligence (AAAI).*

Yue Y.S. and Joachims T. 2008, "Predicting Diverse Subsets Using Structural SVMs." In proceedings of *the International Conference on Machine Learning (ICML)*.

Zaghloul, W., Lee, S.M., Trimi, S. 2009. "Text classification: neural networks vs support vector machines." *Industrial Management & Data Systems*, Vol. **109**, 5, pp.708-717.

Zhang, W., Tang, X., Yoshida, T. 2007. "Text classification with support vector machine and back propagation neural network." In proceeding of *the 7th international conference on computational science (ICCS '07)*, pp.150-157.