NewsFinder: Automating an AI News Service

Joshua Eckroth, Liang Dong, Reid G. Smith, Bruce G. Buchanan

■ *NewsFinder automates the steps involved in finding, selecting, categorizing, and publishing news stories that meet relevance criteria for the artificial intelligence community. The software combines a broad search of online news sources with topic-specific trained models and heuristics. Since August 2010, the program has been used to operate the AI in the News service that is part of the AAAI AITopics website.*

Selecting a small number of interesting news stories each week 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[1] that automatically collects English-language news stories from the web, categorizes them with trained models, filters out uninteresting and unrelated stories based on a set of heuristics, and publishes them in summarized form as the AI in the News service. AI in the News is part of the AAAI AITopics[2] site described previously (Buchanan, Smith, and Glick 2008).

The goal for NewsFinder is to publish a small, select set of news stories that are of general interest to the AI community and to provide categorization metadata to help readers scan for stories that match their specific interests. A news story may be assigned to a single category or to multiple categories (for example, Robots and Vision). The task is similar to asking Google News[3] to find stories about AI, but differs in one important respect. Google News is driven by readers' queries to find stories containing a set of keywords from thousands of news sources. By contrast, NewsFinder uses multiple predetermined queries and Really Simple Syndication (RSS) feeds to find stories from a preferred set of hand-selected sources, as well as from Google News query searches. Google News can be described as a "pull" service because readers must perform queries in order to find news stories that match their interests. NewsFinder, on the other hand, is a "push" service that selects news stories based on

Copyright © 2012, Association for the Advancement of Artificial Intelligence. All rights reserved. ISSN 0738-4602

predicted reader interest and "pushes" those stories to readers by email. We believe that readers of AI in the News expect higher-quality stories than those delivered by Google News, precisely because NewsFinder "pushes" (emails) its weekly alerts. Although NewsFinder and Google News differ in this way, we nevertheless provide an empirical comparison of the two systems.

NewsFinder crawls the web each week looking for recent AI-related news. For each news story, NewsFinder determines which aspects of AI are the main focus of the story (for example, robots, machine learning, vision, and so on), whether the story is a duplicate of another that NewsFinder knows about, and finally whether the story meets certain publication criteria. The best stories are automatically summarized and sorted. After an administrator approves their distribution, the stories are published on a website, inserted into RSS feeds, and sent to email list subscribers.

Design

NewsFinder has two principle components: a weekly crawling, filtering, and publishing component and an offline training component.

Document Representation

As is common in information retrieval applications, we use Salton and Buckley's weighted term-frequency vector space model (Salton and Buckley 1988), commonly known as tf-idf, to represent each news story. Let $W = \{f_1, f_2, \ldots. f_m\}$ be the complete vocabulary set of the crawled news after stemming and stopword filtering. On average, each story contains about 180 different terms after removing stopwords (for example, common English words like *and*, *or*, *not*).[4] The term frequency vector X_i of news story is defined as

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

$$x_{ji} = \log(tf_{ji} + 1) * \log(n/df_j + 1)),$$

where tf_{ji} denotes the frequency of the term $f_j \in W$ in the news story d_i; df_j denotes the number of stories containing word f_j; and n denotes the total number of news stories in the database. The first component, $\log(tf_{ji} + 1)$, is the contribution of the term frequency (tf). The second component, $\log(n/df_j + 1))$, is the contribution of the inverse document frequency (idf). The vector X_i is normalized to unit Euclidean length.

When we discuss training in later sections, note that the inverse document frequency of each term and the number of news stories in the database (n) do not change during weekly web crawling. Rather, we calculate a story's tf-idf value based on a document database that is not changed until retraining is initiated.

Weekly Crawling, Filtering, and Publishing

Each Sunday morning, NewsFinder crawls the web for news stories and publishes a select few. The process consists of three stages: discover, filter, and publish. These stages are shown graphically in figure 1.

Discover

NewsFinder searches 37 preferred online news sources for interesting news. Examples include BBC using search terms "artificial intelligence" and "robots," CNN's "Tech" feed, Discovery's "Robotics" feed, the *New York Times*'s "Artificial Intelligence" and "Robots" feeds, MIT News "AI/Robotics" feed, *IEEE Spectrum*'s "Robotics" feed, and others.[5] Web page and RSS parsers extract story titles, content, and publication dates from the sources. If the news source has an RSS feed, the titles, content, and publication dates of stories are usually already tagged and can be retrieved directly from the RSS format. NewsFinder also queries Google News to find news stories not found in other sources. Queries include "artificial intelligence," "intelligent agent," "machine learning," and 14 other AI-specific phrases.

If the news story does not come from an RSS feed, an open-source heuristic-driven text extractor[6] automatically extracts the main content of a news story, ignoring advertisements, links to other stories, and other unwanted content often found in web pages. For story categorization, summarization, and other filtering and ranking procedures, we are only interested in obtaining an article's title, publisher (for example, BBC, CNN), publication date, and main text (without links).

In addition to crawling news sources, NewsFinder gathers user-submitted news[7] — user submissions come in the form [web address, publication date] — and extracts the web-page content as with any other news source.

Filter

The discover stage produces on average about 170 stories a week. Most of these stories have nothing to do with AI, although they may use AI-related words. The task in the filter stage is to select only the most relevant stories from this collection.

In a prior version of NewsFinder (Dong, Smith, and Buchanan 2011), we asked readers to rate published stories on a 0–5 scale where zero indicates a story is irrelevant and five means a story is highly relevant. This feedback was used to train a multiclass support vector machine (SVM) that predicted a story's rating. The predicted rating would be considered when filtering stories (stories with ratings in the range of 0–2 would be filtered out). However, we subsequently judged that the predictions

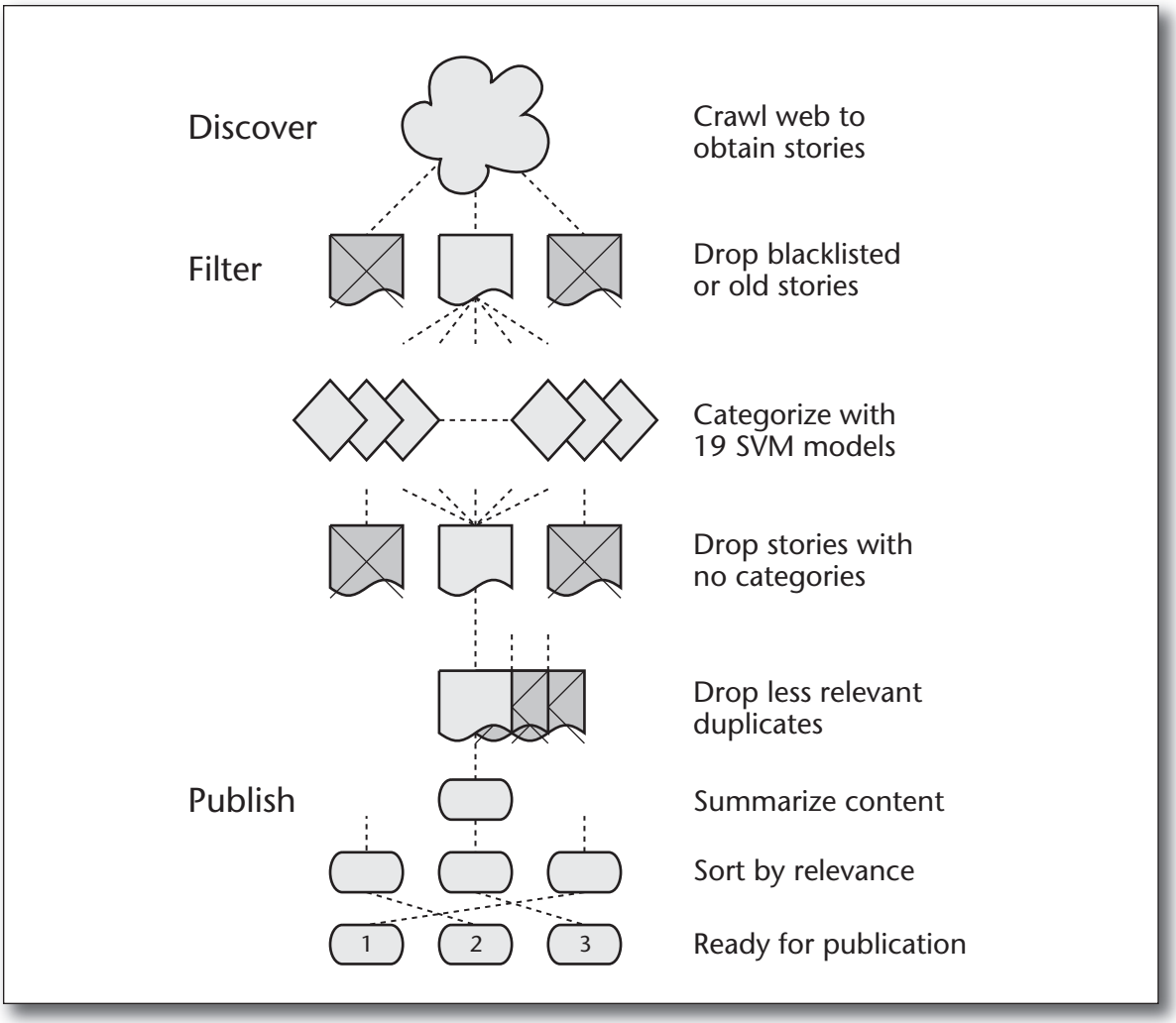

Figure 1 NewsFinder's Three Processing Stages.

Documents are shown as crossed out to indicate that they have been filtered out of the candidate set of news stories. Only candidate stories may be published.

were not sufficiently accurate and abandoned this method of filtering stories. Although we continue to collect and display ratings from readers, we use the ratings for information only and not for training purposes. The filtering process described next seems to be accurate enough for our needs, and has the benefit of greater simplicity.

A news story is first filtered out of the candidate set if it contains any occurrences of profanity or other offensive words on a list known as a blacklist[8] or if its URL matches the URL blacklist (for example, websites that advertise engineering jobs). Next, a news story is filtered out if it does not contain at least one occurrence of each of two different whitelist terms. Subject-matter experts helped us produce the list of 107 terms (the full whitelist is available online[9]), which includes artificial intelligence, Bayes, computer vision, ethical issues, intelligent agents, machine translation, pattern recognition, qualitative reasoning, smart car, and Turing.

Terms without initial capitalization are stemmed before comparing with the news story's content (which is likewise stemmed). We use the Porter stemmer from the Natural Language Toolkit (Loper and Bird 2002).

Next, all occurrences of stopwords are removed from the story's content and the tf-idf vector is constructed. This vector is passed off to the 19 SVMs that individually predict category membership. Stories without any predicted categories are filtered out.

Duplicates are then detected based on a trained similarity threshold. One benefit of representing documents as tf-idf vectors is that we have access to a simple similarity measure, known as the cosine similarity. The dot-product of two document vectors indicates the cosine similarity of the

"Monkeys Control Virtual Limbs With Their Minds"
Date: Oct 5, 2011
Publisher: *Wired*
Categories: Cognitive Science, Robots

Summary: "Now, by implanting electrodes into both the motor and the sensory areas of the brain, researchers have created a virtual prosthetic hand that monkeys control using only their minds, and that enables them to feel virtual textures. Using the first set, the monkey could control a virtual monkey arm on a computer screen and sweep the hand over virtual disks with different textures. By giving the monkey rewards when it identified the right texture, the researchers discovered that it took as few as four training sessions for the animal to consistently distinguish the textures from one another, even when the researchers switched the order of the visually identical disks on the screen. Although the monkeys are all adults, the motor and sensory regions of their brains are amazingly plastic, Nicolelis says: the combination of seeing an appendage that they control and feeling a physical touch tricks them into thinking that the virtual appendage is their own within minutes."

http://www.wired.com/wiredscience/2011/10/monkeys-control-virtual-limbs/

This story was selected for publication due to the following reasons.
Two whitelisted (stemmed) phrases, "brain control" and "virtual prosthet," were found in the text; two categories were selected; and though the story was considered a duplicate of four other stories, this version came from a preferred source *(Wired)*

Figure 2. Example of NewsFinder Output.

two documents; this similarity measure is between 0.0 and 1.0. A similarity threshold (currently 0.17) is used as a cutoff point: if the cosine similarity of two news stories is greater than or equal to this threshold, the stories are labeled as duplicates.

Among a set of duplicate stories, one must be chosen (which may or may not ultimately be published by NewsFinder depending on the number and quality of other candidate stories). We use the following heuristics to choose the top story among duplicates: (1) if one of the duplicates has already been published by NewsFinder in the past 14 days, then do not publish any others; (2) if one of the duplicates has been submitted by a user, then choose the user-submitted story; (3) otherwise, choose the story that comes from the most preferred source (online news sources have been manually ranked for preference[10]), or if two stories come from the same source or both come from Google News searches (whose true sources may not have been seen before, and thus not ranked for preference), then choose the story that has the most predicted categories. We assume that a story with more predicted categories than a duplicate story will take a wider and therefore more interesting view of the news. Finally, if two duplicate stories come from the same source (or Google News) and have the same number of predicted categories, then choose arbitrarily.

What remains at this point is a collection of candidate stories that are considered relevant and nonduplicates. The final task is to further cull this set to produce a small yet diverse selection of stories that represent the week's AI-related news.

Publish

For aesthetic reasons, we have decided that only 12 or fewer stories will be published per week, and ideally no single category (such as robots or machine learning) will have more than eight representatives among the 12 stories. The candidate stories are first sorted by duplicate count (stories with more duplicates are presumed to be more interesting), then sorted by news source credibility, and further sorted by category count (having more categories indicates greater interestingness). Stories are selected in this order until 12 stories have been selected. A story may be skipped if all of its categories are "maxed out," meaning the category already has eight representatives. We only require that one of the story's categories not be "maxed out," so a category may have more than eight representatives in the final published set of stories.

Once 12 stories have been selected, they are summarized using the Open Text Summarizer (Yatsko and Vishnyakov 2007).[11] The summaries (see example in figure 2) are published to an electronic mailing list,[12] on the web,[13] and in 19 category-specific RSS feeds plus one aggregate RSS feed.[14]

Training

Two components of NewsFinder are retrained periodically. The first is the Categorizer; the other is the Duplicate Story Detector.

Categorizer Training

NewsFinder categorizes a story in one or more of the 19 categories shown on the AITopics website: AI overview, agents, applications, cognitive science, education, ethics, games, history, interfaces, machine learning, natural language, philosophy, reasoning, representation, robots, science fiction, speech, systems, and vision.

In an earlier version of NewsFinder (Dong, Smith, and Buchanan 2011), we trained 19 separate centroids (a centroid is a normalized sum of tf-idf document vectors of the members of some category), and predicted that a news story was a member of a category if the cosine similarity (dot-product) of that category centroid and the story's document vector X_i surpassed a threshold. This centroid classification is similar to that described by Han and Karypis (2000) and commonly found in the information retrieval literature. However, it performed poorly on our corpus. We found that our 19 centroids were not well separated, likely due to the fact that news stories naturally fall into more than one category and the categories themselves are not widely separated.

Thus, we changed our approach and trained a separate SVM for each of the 19 categories.[15] Support vector machines, widely used for supervised learning for classification, regression, or other tasks, construct a hyperplane or set of hyperplanes in a high-dimensional space (Burges 1998). They have been applied with success in information-retrieval problems particular to text classification. Recent studies (Dumais et al. 1998; Joachims 1998; Zhang, Tang, and Yoshida 2007; Zaghloul, Lee, and Trimi 2009) have shown that SVMs lead to better text classification than BPNN, Bayes, Rocchio, C4.5, and kNN techniques.

Each SVM is trained on positive and negative examples from the corpus of saved stories and makes a binary prediction for a single category. We do not limit the number of predicted categories for a story, though generally a story matches 1–4 categories. Validation of our categorization procedure is presented later.

Duplicate Story Detector Training

Some events in the world of AI are reported by several news sources. For example, recently IBM showcased its SyNAPSE chips, which are designed to mimic some aspects of a brain. NewsFinder found six news articles, published over a four-day period, which contained essentially the same information.[16] Clearly, it is important to publish, at most, only one story among the duplicates.

NewsFinder's duplicate detection is based on the cosine similarity measure of document vectors. The training procedure involves finding a global threshold that minimizes false positives and maximizes true positives. If two stories have a cosine similarity greater than or equal to this threshold, and the stories were published within two weeks of each other, then they are considered duplicates.

Alternative techniques for detecting duplicates include the shingling algorithm (Broder et al. 1997) and locality sensitive hashing (Charikar 2002). (See Kumar and Govindarajulu [2009] for recent progress in this field.) We chose to use a tf-idf document vector model for its simplicity and speed. Validation of this technique is presented later.

Administration

We recognized early in the development of NewsFinder that it is wise to maintain editorial oversight so that the software does not publish offensive or irrelevant stories. Given that the software obtains its inputs from the web, whose content is unpredictable, there is a possibility that a web page meets NewsFinder's filtering and publishing criteria but is nonetheless outside the scope of AI in the News. Thus, we built the AINewsAdmin interface (figure 3), which allows administrators to remove a news story or modify a story's title, publication date, categorization, and summary in case these fields were erroneously extracted from the story. These modifications can be performed before the

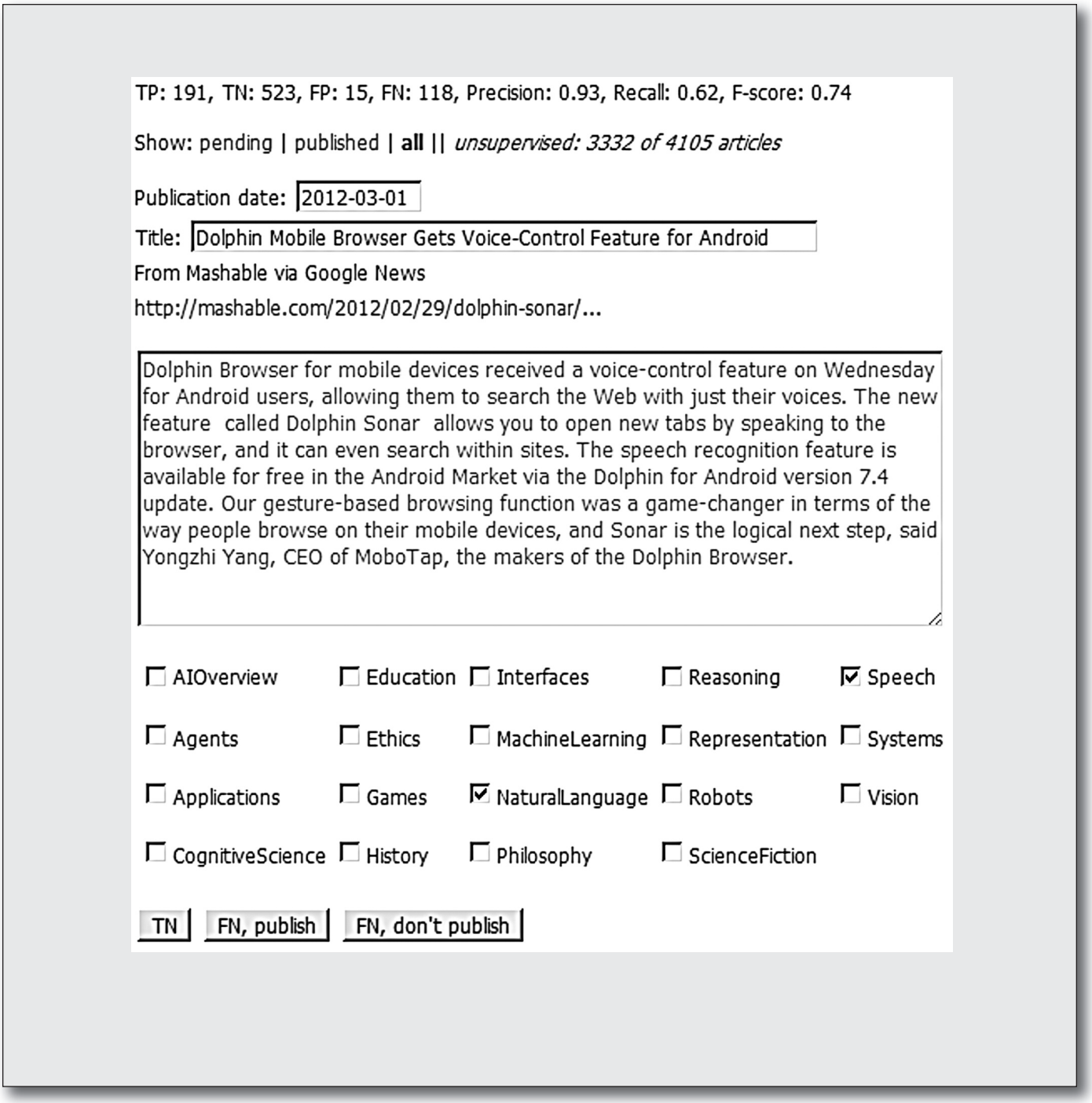

Figure 3. AINewsAdmin Interface.

story is posted in AI in the News or distributed in the email alert.

The interface also enables administrators to add positive and negative examples to the categorization training corpus. The training corpus contains stories deemed relevant to AI in the News readers. For each category, positive examples are those stories that have been tagged with that category by an administrator; negative examples are stories that do not have that category's tag. We believe that any person who is knowledgeable about a subset of our 19 categories can determine if a story is related to categories in that subset. Adding training examples to the corpus need not require any understanding of automated text classification.

In addition to adding examples to the training corpus, an administrator can choose to mark a story as irrelevant and thereby ensure it is kept out of the training corpus.

Not all administrative tasks have been captured by the AINewsAdmin interface. The word blacklist and the URL blacklist may need to be updated. As additional credible news sources are identified, they need to be added to the set of crawled sources. And publication criteria, such as how many stories to publish per category, may need to be modified if too many stories are seen to appear in one category. Each of these modifications can only be performed by editing configuration files or program code, and the modifications are not yet possible through an efficient online interface (see the sidebar "There's More Than AI").

There's More Than AI

The raison d'être of the Innovative Applications of Artificial Intelligence (IAAI) conference is to highlight deployed applications whose value depends on the use of AI technology. NewsFinder fits this bill: the service has been active for several months and depends on several AI components in order to do its job.

However, what is especially interesting about deployed AI applications is that significant effort, if not the majority of effort, is often spent developing and maintaining the non-AI parts of the system. NewsFinder is no different; much human intelligence is required on a regular basis in order to maintain the system and keep the AI components functioning.

NewsFinder requires a variety of parsers and filters in order to extract useful content (such as a news story's text) and to keep out spam or otherwise unwanted content (such as links to other stories, comments, copyright notices, and so on). As of yet, there is no completely automated system that can consistently extract useful content from the web. Websites change formats, and heuristic text extractors sometimes fail to distinguish a story's main text from unmoderated reader comments.

Thus, each week the NewsFinder administrators study the performance of the software and consider modifying the word blacklist, the URL blacklist, the whitelist, and the news sources and search terms. They also determine if parsers must be updated in response to source website redesigns. Before the AI components of NewsFinder are retrained, administrators also have the option to "clean up" the news stories by modifying their content, titles, or categories.

Even though NewsFinder is clearly an application of AI techniques, human intelligence remains a critical component of the overall system.

Validation, Use, and Payoff

As described previously, NewsFinder utilizes 19 trained support vector machines that model the 19 AITopics categories, a trained numeric threshold used by the duplication detector, and heuristic publication criteria. We validate each subsystem in turn, then show how widely AI in the News is read and discuss the payoff of developing the NewsFinder software.

Support Vector Machines for Categorization

The task of choosing one or more categories for a news story is known as multilabel classification. A simple approach to this problem is to build a separate model for each label. This is our approach as well: we train one support vector machine for each label (category). Additionally, we utilize the feature selection algorithm of Chen and Lin (2006) in order to reduce the feature space for the SVM models. Our feature space consists of (stemmed) words found in news stories. Each SVM is trained from about 3000 examples (mostly negative examples), which results in about 37,000 distinct words or features. We determined that most of these words contribute virtually no information about the category membership of a news story, so one may find little benefit in using all 37,000 words during SVM training (and removing a majority of the words from the training data may reduce overfitting). Additionally, training an SVM requires significant computational resources (one of the major drawbacks of an SVM model), and limiting the feature space reduces this cost.

NewsFinder's feature selection algorithm computes the F-score[17] of each feature (stemmed word) and then sorts the features according to this F-score. With experimentation we found that training with only the top 9000 features (about one-quarter of the unique stemmed words found in the corpus) performed best across all categories, and is considerably faster.

All 19 SVMs are trained in this way, independently of each other. For validation, we trained each SVM on 10 percent, 50 percent, and 90 percent of 2940 news stories from the past 10 years (which were categorized by a human). We tested on a nonoverlapping 10 percent of these stories. Additionally, the results (shown in table 1) are averaged over three random training/testing subsets.

The results show that the SVM method of binary category membership is highly accurate in most categories. Less accurate is the applications category, likely because a large number of articles in the corpus are labeled with that category (among oth-

Category (SVM)	Average Accuracy for Percentage of Corpus		
	10%	10%	90 %
AI Overview	88.3%	89.3%	90.0%
Agents	91.3	91.3	92.7
Applications	76.7	76.0	78.0
Cognitive Science	93.7	94.3	94.0
Education	96.0	96.3	98.0
Ethics	88.3	88.0	90.0
Games	89.3	94.7	96.3
History	90.0	91.3	90.7
Interfaces	90.7	90.7	91.0
Machine Learning	85.7	87.3	88.3
Natural Language	91.0	92.0	91.3
Philosophy	97.3	97.7	97.3
Reasoning	96.7	95.7	95.0
Representation	97.3	97.7	97.7
Robots	86.3	89.3	90.0
Science Fiction	93.3	92.3	94.0
Speech	93.7	95.7	95.3
Systems	95.7	95.7	95.0
Vision	90.0	92.7	93.7

Table 1. SVM Accuracy Scores Per Category.

The feature size was fixed at 9000 (details are in the accompanying text). Results are shown for training on 10 percent, 50 percent, 90 percent of the corpus. Each cell reports an average of training accuracies (percent) across three random subsets of the corpus.

True Positive	False Positive	Precision	Recall	F1[18]
992	207	0.827	0.712	0.766

Table 2. Performance of Duplicate Detection.

As judged by one administrator, with threshold equal to 0.17.

True Positive	True Negative	False Positive	False Negative	Precision	Recall	F1
191	523	15	118	0.93	0.62	0.74

Table 3. Performance of NewsFinder's Publication Criteria.

"Interesting and relevant" stories (that is, stories that should be published) are considered positive examples. Stories under consideration range from August 21, 2011, to March 06, 2012 ($N = 847$).

er categories), in part due to reporters' tendency to write about present or future applications.

Duplicate Story Detector

As described previously, the duplicate story detector considers one story to be a duplicate of another if the cosine similarity of the stories exceeds a threshold. Recall that stories are represented as tf-idf vectors; because the tf-idf measure is always nonnegative, the cosine similarity, which is simply the dot product of the document vectors, is always between 0.0 and 1.0. Thus, the threshold that determines whether two stories are duplicates is also between 0.0 and 1.0.

One hundred thresholds between 0.01 and 1.0 (inclusive) were evaluated on a corpus of 1173 news stories, 489 of which have been manually labeled by an administrator as duplicates of one or more others (resulting in 1392 duplicate pairs). The precision, recall, and F1 score resulting from testing with each threshold were measured, and the threshold that maximizes the F1 score was ultimately chosen. In our experiments, this threshold is 0.17. Table 2 shows the performance for this threshold.

Publishing Criteria

To ensure that only interesting and relevant news stories are published by NewsFinder, the heuristics described previously are employed to find the best stories among the large set obtained each week by the automated crawlers. Here we evaluate the performance of these heuristics. We define "true positive" stories as those that NewsFinder published and that a human administrator indicated as interesting and relevant. "False positive" stories were published by NewsFinder but an administrator indicated they should not have been published. "True negative" stories were not published and an administrator agreed and "false negative" stories were not published but an administrator indicated that they should have been. (Note that "false negative" stories are only those that NewsFinder crawled but failed to publish.) Table 3 shows the performance of the publication criteria. NewsFinder achieves high precision but relatively low recall. However, because many readers receive the news in their email inbox, we feel that high precision (fewer irrelevant stories) is more important than high recall (more stories each week).

Use

NewsFinder publishes its weekly news on the AITopics website, in 1 aggregate and 19 topic-specific RSS feeds, and through the AI-Alert electronic mailing list. Usage statistics for these channels show that we are reaching many readers. The

Query	Top story on Google News	Source
Artificial Intelligence	*iOS 5 Voice Assistant to Be "World Changing"	*InformationWeek*
Autonomous Vehicle	Are Autonomous Vehicles Bad News for Fleets?	Hitachi Capital Vehicle Solutions
Cognitive Science	Key to Greatness Is Working Memory, Not Practice	PsychCentral.com
Computer Vision	Commentary: Steve Jobs's Remarkable Vision and Legacy	*Kansas City Star*
Image Understanding	European Space Agency Selects First Two "Cosmic Vision" Missions	Gizmag
Intelligent Agent	*Apple's Siri Is the Fulfillment of a Dream from 1987	Telegraph.co.uk
Machine Learning	Kuzman Ganchev: John Atanasoff Award Is Great Honor to Me	Focus News
Machine Translation	*Bing Powers New Facebook Page Post Translation Tool	Inside Facebook
Neural Network	*Elliot Wave: Extracting Extremes and Predicting Next One by Neural Network	Technorati
Pattern Recognition	Pattern Recognition: How to Find and Photograph Interesting Patterns	Tecca
Robots	Movie Review: "Real Steel	*Los Angeles Times*
Science Fiction	It Has Robots, but Is Real Steel Real Science Fiction?	*Wired News*
Speech Recognition	*Apple Gets Edge with Use of Voice-Recognition Technology	*USA Today*

Table 4. Top News Item from a Subset of Google News Searches.

The searches were performed by NewsFinder (October 6, 2011, 9:30 P.M. UTC). Stories we consider interesting and relevant to the AI community are marked with an asterisk.

AITopics website had a total of 20,945 unique visitors in the 30 days ending on March 6, 2012; the AINews page in particular (which is where NewsFinder publishes its news) had 879 unique visitors in the same time period. The aggregate RSS feed had 185 subscribers as of March 6, 2012. Finally, the AI-Alert weekly email was received by 916 inboxes.

At this time, we are soliciting more feedback from readers in the form of ratings, to be used in future training. We ask in the email and on the website that users rate each news story on a scale of 0–5, where 0 means the story is irrelevant to AI, and 5 means the story is highly relevant. We have also installed Facebook "Recommend" and Twitter "Tweet" buttons on news items to further engage our audience.

Payoff

The cost of developing the program was student stipends for two and a half summers, plus volunteered time by two senior AI scientists. All software libraries used in the development of NewsFinder and the associated AITopics website are free and open source. Maintenance is overseen by volunteers. However, programmers must be hired to consult on major problems when they arise. Before the AI in the News service was automated by NewsFinder, about 10 hours a week was required for a webmaster to find, categorize, summarize, and publish AI-related news. The savings introduced by automating this service offsets the 2.5 student stipends in a 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,[19] which finds stories published

Categories (Chosen by NewsFinder)	Title	Source
Natural Language, Speech	Siri for iPhone 4S	iProgrammer through Google News
Cognitive Seience, Interfaces, Robots	Monkeys Use Brain Interface to Move and Feel Virtual Objects	*IEEE Spectrum*
Robots	Spin-Based Magnetologic Gate to Replace Silicon Chips	KurzweilAI
Machine Learning	Software to Prevent Abuse at the Click of a Mouse	PhysOrg.com through Google News
Interfaces, Robots	Ready for the Robot Revolution?	BBC News through Google News
Speech	Has the Last Fence Fallen? Outperforming Human Emotional Sensitivity	Innovation Investment Journal through Google News

Table 5. NewsFinder's Complete Published News, Ordered by NewsFinder's Ranking.

(October 6, 2011, 9:30 P.M. UTC). Note that every story in this table was categorized as "applications" in addition to the indicated categories; we hide the "applications" category label in published news because most published stories each week match that category.

within the last 30 days. These are selected by query-driven search and ranked with a proprietary algorithm. Because of its widespread use, we compare NewsFinder with this service.

Although the actual list of Google News sources is not published, the stated information from Google is that it crawls more than 4500 English-language sources[20] and altogether more than 25,000 sources in 19 languages around the world.[21] Though the news sources are manually selected by human editors, Google News employs a news ranking algorithm that considers user clicks (popularity), the estimated authority of a publication in a particular topic (credibility), freshness, and geography to determine which stories to show.[22]

Because Google News obtains news stories from many sources, NewsFinder actually performs several searches on Google News in order to find news stories beyond those found in our 37 human-selected sources. Including Google News as a resource enhances the breadth of AI in the News.

NewsFinder does not simply republish Google News stories. As with all stories NewsFinder obtains, Google News stories are filtered according to topic-specific heuristics, then categorized and checked for duplication, and finally summarized and ranked. Google News performs duplicate detection (though we cannot use their algorithms on our own crawled stories). The two systems do not use comparable ranking heuristics or comparable categories for stories beyond high-level categories (technology, entertainment, science, and others.).

To illustrate differences between Google News and NewsFinder, we conducted a sample of the Google News searches that NewsFinder consults, and picked out the top result for each search. (Blogs and press releases, as indicated by Google News, were skipped, as were news stories older than seven days, since blogs, press releases, and old stories are also skipped by NewsFinder.) This experiment is an attempt to simulate a user who is watching several news feeds related to AI. Because Google News searches are user-initiated, one may imagine that readers of Google News are more forgiving if search results are irrelevant or uninteresting. NewsFinder, on the other hand, delivers the news in readers' inboxes (among other outlets), and thus must maintain relatively high precision. The AINewsAdmin interface allows human oversight of NewsFinder's output, while Google News is completely automated.

Results from searches on Google News are shown in table 4. NewsFinder was executed at the same time the Google News searches were performed; its output is shown in table 5. The time and date of the experiment are arbitrary.

We see that Google News returns some irrelevant stories such as "Commentary: Steve Jobs's Remarkable Vision and Legacy"[23] for the "computer vision" query. On the other hand, NewsFinder missed some stories that we consider interesting and relevant to AI that Google News did find. One case is the story "Bing Powers New Facebook Page Post Translation Tool,"[24] which describes how machine translation will be further integrated into Facebook, allowing users to communicate with each other across disparate languages. NewsFinder

did crawl this story but only found the whitelisted term "machine translation," and no other whitelisted terms. NewsFinder's publication criteria require that at least two distinct whitelisted terms be found in a story. An administrator may choose to alter the whitelist so that stories such as this one are published in the future; however, modifying the whitelist or other heuristics affects NewsFinder's overall accuracy, and this impact is often unpredictable.

We believe that the comparison of Google News search results with (unmodified) NewsFinder output shows that NewsFinder, and the AI in the News service, has a higher signal-to-noise ratio than a collection of query-driven news feeds. Google News's lower signal-to-noise ratio may be acceptable to a reader who wants the flexibility of searching for news through specific queries. NewsFinder, on the other hand, is not flexible in this way. However, because it is designed specifically to find AI-related news, the quality of news published by NewsFinder generally exceeds that of a collection of Google News searches.

Conclusions and Future Work

Replacing a time-consuming, yet intelligent, manual operation with an AI-enhanced system is a natural project for AAAI. The AI in the News service was the subject of this exact line of thought, and the success of the NewsFinder system demonstrates that automating such a task is both feasible and advantageous.

Several design decisions have made NewsFinder an effective yet relatively inexpensive AI project. We used open source and free software for important features like support vector machine training, natural language processing, data storage, web crawling, and website content management. Several heuristics and human oversight are built into the system to ensure high-quality output. Additionally, because NewsFinder attempts to cast a broad net, we implemented strict filters so that it can provide information that is both relevant and extensive.

Finally, although NewsFinder presently seeks out AI-related news, there is no aspect of its design that intentionally restricts its domain to AI. Prima facie, the requirements to apply the system to another domain are: subject matter experts to serve as administrators; a set of topics that characterize the domain; a corpus of news stories to train the topic-specific support vector machines; and a whitelist, blacklists, and source-specific parsers. However, one might expect to encounter hidden interdependencies of the heuristics and other components of NewsFinder that would initially complicate adaptation to a new domain.

The system should continue to improve over time as additional training examples are added to its corpus, but work remains to be done to make more of NewsFinder accessible to automated learning. Other improvements we would like to pursue include more useful analytical tools such as the ability to determine how strongly whitelist words are correlated to category membership, how category membership relates to news sources (for example, do some news sources always deliver robot stories?), and whether certain categories are "trending" over time (for example, is cognitive science becoming a hot news topic and, if so, why).

NewsFinder was constructed for the Association for the Advancement of Artificial Intelligence from open-source modules at low cost and supplies an effective service for the AI community.

Acknowledgments

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

Notes

1. NewsFinder's code is available at github.com/AAAI.
2. See aaai.org/AITopics/AINews.
3. See news.google.com.
4. We use a modified version of the MIT list of 583 terms, available at jmlr.csail.mit.edu/papers/volume5/lewis04a.
5. The full list of sources can be found at aaai.org/AITopics/NewsSources.
6. See Justext, code.google.com/p/justext.
7. News may be submitted at www.aaai.org/AITopics/SubmitNewContent.
8. We use a modification of this blacklist. See urbanoalvarez.es/blog/2008/04/04/bad-words-list/.
9. See github.com/AAAI/AINewsSupplementary/blob/master/resource/whitelist.txt.
10. See www.aaai.org/AITopics/NewsSources for the full list of sources and their preference ratings.
11. Available at libots.sourceforge.net.
12. See www.aaai.org/AITopics/mailinglist. To prevent spam, the AAAI distribution-list server does not allow automatic sending of email. Thus a member of the AAAI staff must manually send a prepared message to subscribers each week.
13. See www.aaai.org/AITopics/AINews.
14. See www.aaai.org/AITopics/AINews#feeds.
15. We use the LibSVM open source library (Chang and Lin 2011) for its extensive programming language support and speed.
16. Visit www.aaai.org/AITopics/AIArticles/2011-2669 to see the list of duplicate SyNAPSE stories found by NewsFinder.
17. A feature with a larger F-score is more discriminative, with respect to a certain binary SVM. Details can be found in Chen and Lin (2006).
18. Precision = TruePos/(TruePos+FalsePos). Recall = TruePos/(TruePos+FalseNeg). F1 = 2*(Precision*Recall)/(Preci-

sion+Recall); note that this F1 score is not related to the F-score used to select features for the SVMs.

19. See news.google.com.

20. See en.wikipedia.org/wiki/Google_News.

21. See googlenewsblog.blogspot.com/2009/12/same-protocol-more-options-for-news.html.

22. See searchengineland.com/google-news-ranking-stories-30424.

23. See www.kansascity.com/2011/10/06/3190839/commentary-steve-jobs-remarkable.html.

24. See www.insidefacebook.com/2011/10/05/bing-translate-pages.

References

Broder, A. Z.; Glassman, S. C.; Manasse, M. S.; and Zweig, G. 1997. Syntactic Clustering of the Web. *Computer Networks and ISDN Systems* 29(8–13): 1157–1166.

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

Burges, C. J. C. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2(2): 121–167.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* (TIST) 2(3): 27–53.

Charikar, M. S. 2002. Similarity Estimation Techniques from Rounding Algorithms. In *Proceedings of the Thirty-fourth Annual ACM Symposium on Theory of Computing,* 380–388. New York: Association for Computing Machinery.

Chen, Y.-W., and Lin, C.-J. 2006. Combining SVMs with Various Feature Selection Strategies. *Feature Extraction: Studies in Fuzziness and Soft Computing* 207: 315–324.

Dong, L.; Smith, R. G.; and Buchanan, B. G. 2011. NewsFinder: Automating an Artificial Intelligence News Service. In *Proceedings of the Twenty-Third IAAI Conference on Innovative Applications of Artificial Intelligence* (IAAI 11). Palo Alto, CA: AAAI Press.

Dumais, S.; Platt, J.; Heckerman, D.; and Sahami, M. 1998. Inductive Learning Algorithms and Representations for Text Categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management* (CIKM 98), 148–155. New York: Association for Computing Machinery.

Han, E.-H. S., and Karypis, G. 2000. Centroid-Based Document Classification: Analysis and Experimental Results. In *Principles of Data Mining and Knowledge Discovery,* 116–123. Berlin: Springer.

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning,* 137–142. Berlin: Springer.

Kumar, J. P., and Govindarajulu, P. 2009. Duplicate and Near Duplicate Documents 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,* 63–70. Philadelphia, PA: Association for Computational Linguistics.

Salton, G., and Buckley, C. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management* 24(5): 513–523.

Yatsko, V. A., and Vishnyakov T. N. 2007. A Method for Evaluating Modern Systems of Automatic Text Summarization. *Automatic Documentation and Mathematical Linguistics* 41(3): 93–103.

Zaghloul, W.; Lee, S. M.; and Trimi, S. 2009. Text Classification: Neural Networks Versus Support Vector Machines. *Industrial Management and Data Systems* 109(5): 708–717.

Zhang, W.; Tang, X.; and Yoshida, T. 2007. Text Classification with Support Vector Machine and Back Propagation Neural Network. In *Proceedings of the Seventh International Conference on Computational Science* (ICCS 07), 150–157. Berlin: Springer.

Joshua Eckroth is a Ph.D. student at the Ohio State University. He is studying artificial intelligence, cognitive science, and logic. His research focuses on how an abductive reasoning agent may employ novel metareasoning strategies to improve its understanding of a changing world.

Liang Dong is a Ph.D. candidate in computer science at the School of Computing of Clemson University, South Carolina. His research interests lie in search engines and three-dimensional animation. He is expecting to graduate in May 2012. He obtained his M.S. and B.S. in Shanghai Jiao Tong University in 2007 and 2004, respectively.

Reid G. Smith is the enterprise content management director and IT upstream services manager of Marathon Oil Corporation. Prior to joining Marathon, he was responsible for information solutions at Medstory. Earlier, he led the knowledge management program at Schlumberger and managed a number of the company's research laboratories in the United States and the United Kingdom. Smith holds a Ph.D. in electrical engineering from Stanford University and is a fellow of the Association for the Advancement of Artificial Intelligence.

Bruce G. Buchanan is the University Professor of Computer Science Emeritus at the University of Pittsburgh. Before joining the Pitt faculty as a professor of computer science, medicine, and philosophy, he was a professor of computer science (research) at Stanford where he worked on the Dendral, Meta-Dendral, Mycin, and Protean systems. He has supervised more than two dozen Ph.D. dissertations in AI and related fields. Buchanan holds a Ph.D. in philosophy from Michigan State University; he is a fellow of AAAI and the American College of Medical Informatics, an elected member of the National Academy of Science Institute of Medicine, and has served as the secretary-treasurer and president of AAAI.