

# NewsFinder: Automating an AI News Service

---

2013012568 신채은  
(SHIN CHEEUN)  
2013012654 이다연  
(LEE DAYOUN)

# INDEX

---

**01** What is the NewsFinder

**02** Research Background

**03** Research Goal

**04** Process

**05** Crawling, filtering , publishing

**06** Training

**07** Validation

**08** Contribution

**09** Comparison with Google News

**10** Lesson Learned

---

# » What is the NewsFinder?

The software combines a broad search of online news sources with topic-specific trained models and heuristics.

The program has been used to operate the AI in the News service that is part of the AAAI AITopics Website



## Topics

- ▶ [AI Overview](#)
- ▶ [Applications](#)
- ▶ [Cognitive Science](#)
- ▶ [Education](#)
- ▶ [Ethics & Social Issues](#)
- ▶ [Games & Puzzles](#)
- ▶ [History](#)
- ▶ [Interfaces](#)
- ▶ [Machine Learning](#)
- ▶ [Natural Language](#)
- ▶ [Philosophy](#)
- ▶ [Representation & Reasoning](#)
- ▶ [Robots](#)
- ▶ [Science Fiction](#)
- ▶ [Speech](#)
- ▶ [Systems & Languages](#)
- ▶ [Vision](#)
- ▶ [Web & AI](#)

## Goals & Visions

*AI is thought to be an impossible dream by many. But not to us in AI. It is not only a possible dream, but, from one point of view, AI has been a reality that has been demonstrating results for nearly 40 years. And the future promises to generate an impact greater by orders of magnitude than progress to date.*

*- Raj Reddy, "To Dream the Possible Dream"*

Image from [Kathryn McCallum](#) on Flickr.



[AI Overview](#)

## Contents

### Good Starting Places

### General Readings

### News

## General Readings

*AI in the News*

[Why Can't My Computer Understand Me?](#)

- Representation & Reasoning
- Robots
- Science Fiction
- Speech
- Systems & Languages
- Vision
- Web & AI

## Collections

- Interviews
- Tributes
- News
- Classics
- Podcasts
- Videos
- Course materials
- FAQs

## About Us

- About AITopics
- Project Notes
- Editorial Policies
- Privacy Policy
- Terms of Use
- Contact Us
- Login

## Contents

### Good Starting Places

#### General Readings

#### News

## Good Starting Places

### AI in the News

## The Myth Of AI: A Conversation with Jaron Lanier



*If you talk about AI as a set of techniques, as a field of study in mathematics or engineering, it brings benefits. If we talk about AI as a mythology of creating a post-human species, it creates a series of problems that I've just gone over, which include acceptance of bad user interfaces, where you can't tell if you're being manipulated or not, and everything is ambiguous. It creates incompetence, because you don't know whether recommendations are coming from anything real or just self-fulfilling prophecies from a manipulative system that spun off on its own, and economic negativity, because you're gradually pulling formal economic benefits away from the people who supply the data that makes the scheme work.*

(with video)

Nov 20 2014, By Brockman, John



Jaron  
Lanier



Frankenstein, video

### Publication

## The Gardens of Learning: A Vision for AI

- › Vision
- › Web & AI

## Collections

- › Interviews
- › Tributes
- › News
- › Classics
- › Podcasts
- › Videos
- › Course materials
- › FAQs

## About Us

- › About AITopics
- › Project Notes
- › Editorial Policies
- › Privacy Policy
- › Terms of Use
- › Contact Us
- › Login



## News

### AI in the News

## The Myth Of AI: A Conversation with Jaron Lanier



*If you talk about AI as a set of techniques, as a field of study in mathematics or engineering, it brings benefits. If we talk about AI as a mythology of creating a post-human species, it creates a series of problems that I've just gone over, which include acceptance of bad user interfaces, where you can't tell if you're being manipulated or not, and everything is ambiguous. It creates incompetence, because you don't know whether recommendations are coming from anything real or just self-fulfilling prophecies from a manipulative system that spun off on its own, and economic negativity, because you're gradually pulling formal economic benefits away from the people who supply the data that makes the scheme work.*

(with video)

Nov 20 2014, By Brockman, John



Jaron  
Lanier



Frankenstein, video

### AI in the News

## Why Can't My Computer Understand Me?



Hector Levesque thinks his computer is stupid—and that yours is, too. Siri and Google's voice searches may be able to understand canned sentences like "What movies are showing near me at seven o'clock?" but what about questions—"Can an alligator run the hundred-metre hurdles?"—that nobody has heard before? Any ordinary adult can figure

that one out.



Fri, Oct 02, 2015

CONVERSATIONS

VIDEOS

ANNUAL QUESTION

EVENTS

NEWS

LIBRARY

ABOUT

CONVERSATION : TECHNOLOGY

## The Myth Of AI

A Conversation With **Jaron Lanier** [11.14.14]



The idea that computers are people has a long and storied history. It goes back to the very origins of computers, and even from before. There's always been a question about whether a program is some kind of autonomy at the very least, or it wouldn't be

### WHAT'S RELATED

#### People

**Jaron Lanier**

Computer Scientist; Musician; Author, Who Owns The Future?

#### Contributors

**George Church**Author, Regenesis; Professor, Harvard University...  
[ Read ]**Peter Diamandis**

Chairman/CEO, X PRIZE Foundation; Co-author, Bold [ Read ]

**Lee Smolin**Physicist, Perimeter Institute; Author, Time...  
[ Read ]**Rodney A. Brooks**Robotician; Panasonic Professor of Robotics (...)  
[ Read ]**Nathan Myhrvold**

CEO and Managing Director, Intellectual Ventures... [ Read ]

**George Dyson**

Science Historian; Author, Turing's... [ Read ]





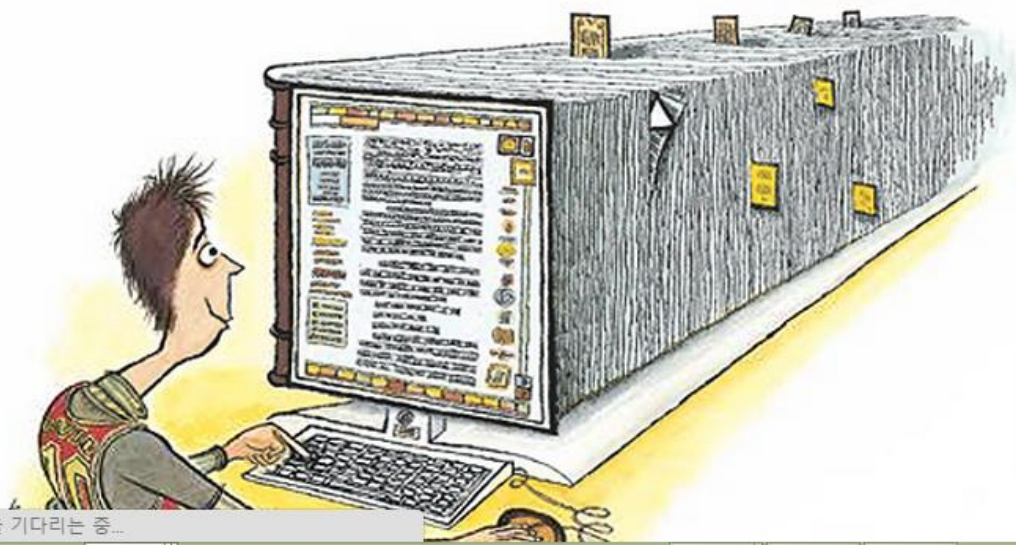
AUGUST 14, 2013

## WHY CAN'T MY COMPUTER UNDERSTAND ME?

BY GARY MARCUS



ELEMENTS



## MOST POPULAR

### 1. What Old Age Is Really Like

BY CERIDWEN DOVEY

### 2. China's Butler Boom

BY BIANCA BOSKER



# AI in the News

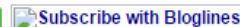
syndicated content powered by FeedBurner

FeedBurner makes it easy to receive content updates in My Yahoo!, Newsgator, Bloglines, and other news readers.

[Learn more about syndication and FeedBurner...](#)

## Subscribe Now!

...with web-based news readers. Click your choice below:



...with other readers:

(Choose Your Reader) ▼

[View Feed XML](#)

## Current Feed Content

### Google introduces Pixel C tablet, new Nexus phones

Posted: Tue, 29 Sep 2015 00:00:00 +0000



Google CEO Sundar Pichai announces a new Pixel C tablet that can be connected to a laptop for desktop style use, in San Francisco on September 29, 2015. (Justin Sullivan/Getty Images) Since September 29, 2015, Google introduced a tablet-laptop hybrid Tuesday, along with two new Nexus phones and a new Chromecast that together marked the company's first big product launch since reorganizing under parent company Alphabet.

San Jose Mercury News - Technology

[Link](#)

RSS feed

## » Research Background

Manually finding and posting stories that are likely to be interesting is **time-consuming**.

Therefore, they have developed an AI program

# » Research Goal

## The Primary goal

to be a trusted, timely, and educational source of AI-related news that is collected, summarized, and categorized nowhere else.

## NewsFinder Mission

- Provide links to relevant and interesting news stories about AI
- Automate the service of finding and summarizing stories
- Implement the service, well-documented program using available tools to provide a foundation for future improvement.

# » Process

Crawling → filtering  
→ training (categorizing)  
→ ranking  
→ publishing.

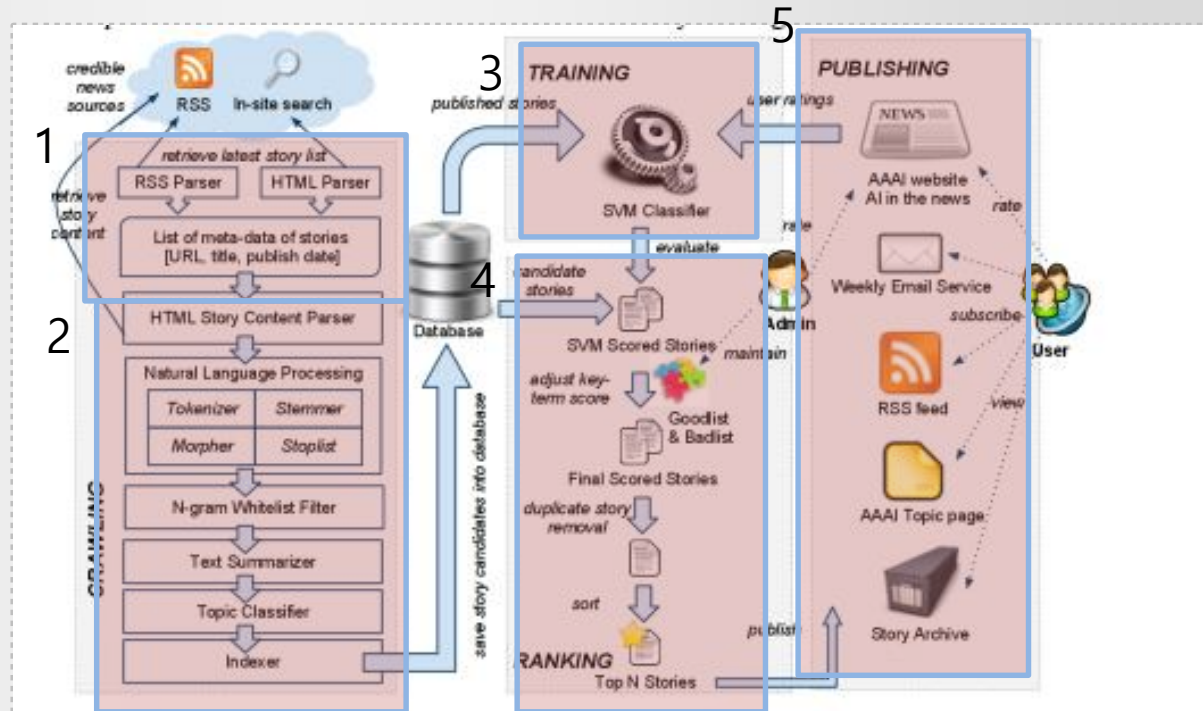


Fig. 1. NewsFinder Procedure Diagram

# » **Crawling** , Filtering, Publishing

- Searches **37 preferred online news sources** for interesting news.

Ex) 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

- Online sources are divided into three groups:

General sources	Technology sources	AI sources
BBC, CNN, Discovery, Forbes, The Guardian, LA Times, MSNBC, New York Times, NPR, USA Today, Wall Street Journal, Washington Post;	CNet, MIT News, New Scientist, Popular Science, Scientific American, Wired, and ZDNet, which <b>focus on science and technology</b>	Kurzweil.Net 2 and Robots.Net 3 , <b>which focus on news about artificial intelligence</b>

# » **Crawling** , Filtering, Publishing

- 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.
- Also queries **Google News** to find news stories not found in other sources.
- an open-source **heuristic-driven text extractor** automatically extracts the main content of a news story, ignoring advertisements, links to other stories, and other unwanted content often found in web pages
- Only interested in obtaining an article's title, publisher , publication date, and main text for process.
- Also gathers **user-submitted news**.

# » Crawling , Filtering, Publishing

1. Filtered out of the candidate set if it contains any occurrences of profanity or other offensive words.
2. A news story is filtered out if it does not contain at least one occurrence of each of two different whitelist terms.  
Ex) artificial intelligence, Bayes, computer vision, ethical issues, intelligent agents, machine translation, pattern recognition .....

**\*Whitelist Terms:** indicate relevance and interest terms in the story.



## » Crawling , Filtering, Publishing

3. Remove Stopwords -> construct **tf-idf** vector -> 19  
**SVMs model**

→ Stories without any predicted categories are filtered out.

\*TF : Term Frequency

\*DF : Document Frequency

\*IDF : Inverse Document Frequency ( $IDF \propto$  importance of word)

## » Crawling , Filtering, Publishing

4. Duplicates are detected according to a trained **similarity threshold**.

→ The dot-product of two document vectors indicates the cosine similarity of the two documents

\*  $0 \leq \text{similarity measure} \leq 1$  \* A similarity Threshold : 0.17 ;  
if *cosine similarity*  $\geq$  *threshold*, then the stories are duplicates

# » Crawling , Filtering, Publishing

- Among a set of duplicate stories, one must be chosen
- Use the following heuristics to choose the top story among duplicates:
  - (1) has already been published by NewsFinder in the past 14 days ✖
  - (2) user-submitted story ○
  - (3) the story that comes from the most preferred source ○
  - (4) if two stories come from the same source or both come from Google News searches ○
- To further cull set(relevant and nonduplicates) to produce a small

# » Crawling , Filtering, Publishing

Only 12 or fewer stories will be published per week,

1. The candidate stories are sorted by

(1)duplicate count ( duplicate count  , interestingness  )

(2)news source credibility

(3)category count ( category count  , interestingness  )

If all of its categories are “maxed out”

→ A story may be skipped

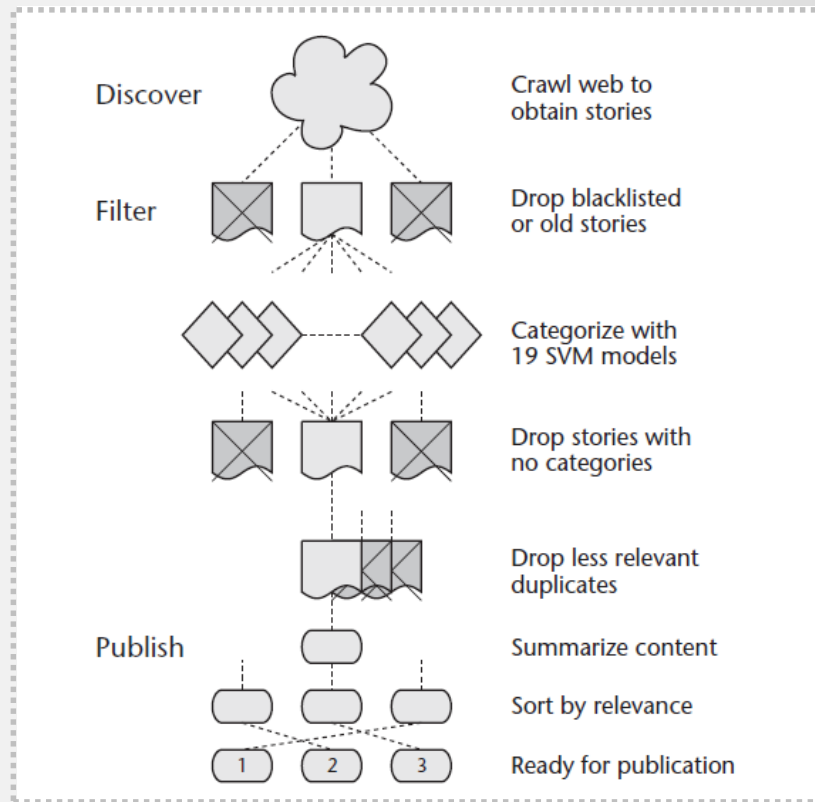
2. summarized using the Open Text Summarizer  
(Yatsko and Vishnyakov 2007).

# » Crawling , Filtering, Publishing

**Publishing** The selected stories are published via five channels

- 01 the Latest AI in the News summary page of **AITopics**
- 02 the weekly “**AIAlert**” **e-mail** message to subscribers10,
- 03 the **RSS feeds** associated with the AITopics major topics
- 04 the **AITopics news archive**
- 05 **individual story pages** on AITopics.

# » Crawling , Filtering, Publishing

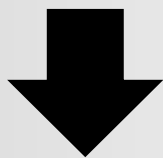


# » Training

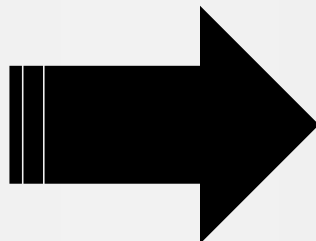
## Categorizer Training (SVM classifier)



the 19 categories



In a earlier version,  
They trained 19 separated  
centroids



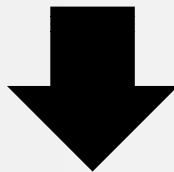
They changed  
approach and trained  
a separated **SVM** for  
each of the 19  
categories

\*A centroid : one of classification that categorizes text through a set of points representing the documents



# » Training Duplicate Story Detector Training

NewsFinder's duplicate detection is based on the cosine similarity of document. The training procedure involves finding a global threshold.



If two stories have a cosine similarity  $\geq$  this threshold, then the stories are considered duplicates.

\*Cosine similarity: measure the similarity between two cosine values when the query and actual documents represented by a vector

\*Threshold: threshold is the minimum value of measure that can be said to be similar.

# » Validation

## SVMs(Support Vector Machines) for Categorization

The task of choosing one or more categories for news story is known as multi-label classification. A simple approach to this problem is to build a separate model for each label.

They trained one support vector machine for each label (category).

So, NewsFinder utilizes 19 trained SVM.

For validation, they trained each SVM on 10%, 50%, 90% of 2940 news stories from the past 10 years.

The results show that the SVM method is highly accurate in most categories.

Category (SVM)	Average Accuracy for Percentage of Corpus		
	10%	50%	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
Educ	96.0	96.3	98.0
Ethics	88.3	88.0	90.0
Games	89.3	94.7	96.3
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

A large number of articles are labeled with application category. Because of reporter's tendency to write about present or future applications.

Table 1. SVM Accuracy Scores Per Category.

# » Validation

## Publishing Criteria

	Relevant	Not relevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

Error!

# » Validation

## Publishing Criteria

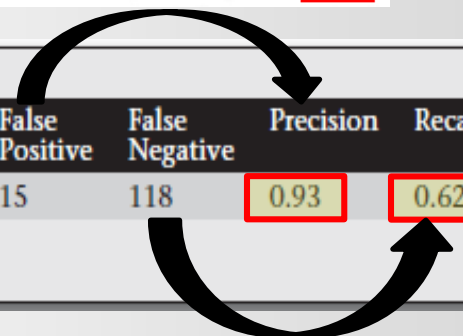
**Precision** = The percentage of relevant documents among retrieval documents.

**Recall** = The percentage of retrieval documents among relevant documents.

NewsFinder achieves high precision but, relatively low recall.

However, (because many readers receive the news in their email inbox) they feel that high precision is more important than high recall.

$$P = tp / (tp + fp)$$



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

$$R = tp / (tp + fn)$$

# » Contribution-Use

AI Topics Website	Aggregate RSS feed	AI-Alert weekly email
20,945 visitors (2012,March)	185 subscribers	916 inboxes
→AINews : 879 visitors		

- more feedback from readers in the form of ratings, to be used in future training.
- installed Facebook “Recommend”  
Twitter “Tweet” buttons on news items  
to further engage our audience.

# » Contribution-Pay Off

- The savings introduced by automating this service offsets the 2.5 student stipends in a year or less.
- Additional benefits: from the consistency , reliability of an automated Service +
- the unquantifiable benefits of providing useful information to the AI community

# » Comparison with Google News

NewsFinder	Google News
<ul style="list-style-type: none"><li>- Use <u>multiple predetermined queries</u> and really simple syndication (RSS)</li><li>- Described '<u>Push</u>' because <u>service selects news stories</u> based on predicted reader interest and 'pushes' those stories to readers by email</li><li>- Perform several searches on Google News in order to find news stories but doesn't simply republish Google News stories.</li></ul>	<ul style="list-style-type: none"><li>- Driven by <u>reader's queries</u> (query-driven search)</li><li>- Described '<u>Pull</u>' because <u>readers must perform queries</u> in order to find news stories that match their interests</li><li>- Crawls more than 4500 English-language sources and more than 25000 sources in 19 languages around world.</li></ul>



# » Comparison with Google News

NewsFinder	Google News
<ul style="list-style-type: none"><li>- <u>Missed some stories</u> that consider <u>interesting and relevant to AI</u> that Google News did find.</li><li>- (ex. Publication criteria require that at least two distinct whitelisted terms but, missed news only is founded just on whitelisted term.</li><li>- Higher signal-to-noise ratio</li><li>- <u>Not flexible</u> but, it is <u>designed specifically to find AI related news</u>, so good quality of news</li></ul>	<ul style="list-style-type: none"><li>- <u>Returns some irrelevant stories.</u></li><li>- Lower signal-to-noise ratio</li><li>- Acceptable to a reader who wants the <u>flexibility</u> of searching for news <u>through specific queries.</u></li></ul>

## » Lesson Learned

- Have experienced **Information Retrieval(IR)** field during the study
- More complex and More Mathematical than expected
- Apply to other fields

THE

END

Thank You