NewsInEssence: A System For Domain-Independent, Real-Time News Clustering and Multi-Document Summarization

Dragomir R. Radev^{*†}, Sasha Blair-Goldensohn^{*}, Zhu Zhang^{*}, Revathi Sundara Raghavan[†] *School of Information [†]Department of EECS University of Michigan Ann Arbor, MI 48109 {radev,sashabg,zhuzhang,rsundara}@umich.edu

1. INTRODUCTION

NEWSINESSENCE is a system for finding, visualizing and summarizing a topic-based cluster of news stories. In the generic scenario for NEWSINESSENCE, a user selects a single news story from a news Web site. Our system then searches other live sources of news for other stories related to the same event and produces summaries of a subset of the stories that it finds, according to parameters specified by the user.

2. THE NEWSINESSENCE SYSTEM

NewsInEssence's search agent, NewsTroll, runs in two phases. First, it looks for related articles by traversing links from the page containing the seed article. Using the seed article and any related articles it finds in this way, the agent then decides on a set of keywords for further search. In the second phase, it attempts to add to the cluster of related articles by going to the search engines of various news websites and using the keywords which it found in the first phase as search terms.

In both phases, NewsTroll selectively follows hyperlinks with the aim of reaching pages which contain related stories and/or further hyperlinks to related stories pages.

Both general and site-specific rules help NewsTroll determine which URLs are likely to be useful. Only if NewsTroll determines that a URL is "interesting", will it go to the Internet to fetch the new page. A more stringent set of rules are applied to determine whether the URL is likely to be a news story itself. If so, the similarity of its text to that of the original seed page is computed using an IDF-weighted vector measure. If the similarity is above a certain threshold, the page is considered to contain a related article and added to the cluster. The user may use our web interface (Figure 2) to adjust the similarity threshold used in a given search.

Using several levels of filtering, NewsTroll is able to screen out large numbers web pages quite efficiently. The expensive operation of testing lexical similarity is reserved for the small number of pages which NewsTroll finds interesting. Consequently, the agent can return useful results in real time.

3. ANNOTATED SAMPLE RUN

The example begins when we find a news article we would like to read more about. In this case we pick a story is about a breaking story regarding one of President-Elect Bush's cabinet nominees (see Figure 1).

We input the URL using the web interface of the NEWSINESSENCE system, then select our search options, click 'Proceed' and wait for our results (see Figure 2).

In response to the user query, NewsTroll begins looking for related articles linked from the chosen start page. In a selection from the agent's output log in Figure 3, we can see that it extracts and tests links from the page, and decides to test one which looks like a news article. We then see that it tests this article and determines it to be related. This article is added to the initial cluster, from which the list of top keywords is drawn.

In its secondary phase, NewsTroll inputs its keywords to the search engines of news sites and lets them do the work of finding stories. Since we have selected good keywords, most of the links seen by NewsTroll in this part of the search are indeed related articles (see Figure 4). Upon exiting, NewsTroll reports the number of links it has considered, followed, tested, and retrieved (see Figure 4).

The system's web interface reports its progress to the user in real time and provides a link to the visualization GUI once the cluster is complete (Figure 5). Using the GUI, the user can select which of the articles to summarize (see Figures 6 and 7). Figure 8 shows the output of the cluster summarizer.

4. FUTURE WORK

We are currently working on the integration of Cross-Document structure theory (CST) [1] with NEWSINESSENCE. CST is used to describe relations between textual units in multi-document clusters. It is used for example to identify which portions of a cluster contain background information, which sections are redundant, and which ones contain additional information about an event.

5. **REFERENCES**

 Dragomir Radev. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In *Proceedings, 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, Hong Kong, October 2000.

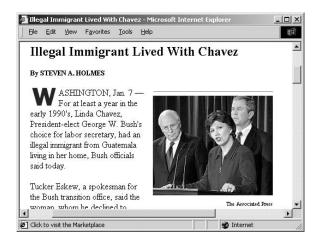


Figure 1: Seed article.

	NewsInE	ssence Demo		
		Story URL		
		2001/01/08/pol/tics/08CHAV h/		
	2	ources		
C all available sources & selected sources on C pages inked to origi		 □ BBC (news bbc couk) □ CINN (www cnn con) □ MSNBC (www minbc) □ NY Times (www nytime □ USA Today (www isab □ V Yahool (dalynews yahoo) 	es.com) oday.com)	
	Search	Thresholds		
Minimum Article Similarity	Search Timeout (min	0	Maximum Sim Testa/Site	
C high (strongly related) @ medium (related) C low (somewhat related)	(if set to none, search w is met.)	il run until max pagre threshold	@ manamum (3) C default (5) C maximum (8)	

Figure 2: User interface.

- · · ·	@svarog.si.umich.edu
	source of seed url for lexical comparisons
	seed page is : Illegal Immigrant Lived With Chavez
	run commencing for seed: www.nytimes.com/2001/01/08/politics/08CHAV.html
nccp://	AND ANGLINES.COM/2001/01/08/POILCICS/06CHHY.ICHI
Extract	links from http://www.nytimes.com/2001/01/08/politics/08CHAV.html
http	://www.nytimes.com/pages/index.html OK
	p://www.nytimes.com/info/contents/siteindex.html OK
	://www.nytimes.com/search/daily/ OK
	://www.nytimes.com/2001/01/08/politics/08CHAV.html?pagewanted=all0 ://www.nytimes.com/2001/01/07/politics/07LAB0.html 0K
	<pre>//www.nytimes.com/library/politics/whouse/whouse-cabinet.html OK</pre>
U/ -/ + /	z page title: The New York Times Site Index
	pattern did not match article comparison pattern for site continuin
GIT	pactern did not watch a citie comparison pactern for site continuin
	g page title: Union Leaders Plan Fight Against Bush's Labor Selectio ning lexsim comparison Passed! (score = 0,74)
	ere the highest scoring words in the similarity calculation:
	vez, eskew, sweeney, ws, bush, labor, affirmative, wage, enforce
cha	
cha Primary	vez, eskew, sweeney, ms, bush, labor, affirmative, wage, enforce
cha Primary Search f Search f	vez, eskew, sweeney, mš, bush, labor, affirmatīve, wage, enforce run ending for seed: ıerm no. 1 would be: chavez (score = 4) cerm no.2 would be: eskew (score = 3)
cha Primary Search 1 Search 1 Search 1 Search 1	<pre>rez, eskew, sweeney, ms, bush, labor, affirmative, wage, enforce run ending for seed: run ending for seed: run ending for sharez (score = 4) erm no. 2 would be: eskew (score = 3) erm no. 2 would be: eskew eney (score = 2,66)</pre>
Char Primary Search 1 Search 1 Search 1 Search 1 Search 1	rez, eskew, sweeney, mŝ, bush, labor, affirmative, wage, enforce run ending for seed: Lerm no. 1 would be: chavez (score = 4) Lerm no. 3 would be: sweeney (score = 2,66) Lerm no. 3 would be: sweeney (score = 2,5)
Char Primary Search 1 Search 1 Search 1 Search 1 Search 1 Search 1	<pre>rez, eskew, sweeney, ms, bush, labor, affirmative, wage, enforce run ending for seed: run ending for seed: rerm no. 2 would be: chavez (score = 4) rerm no. 2 would be: eskew (score = 2,6) rerm no. 4 would be: ms (score = 2,5) rerm no. 4 would be: ms (score = 2,4)</pre>
Char Primary Search 1 Search 1 Search 1 Search 1 Search 1 Search 1	rez, eskew, sweeney, mŝ, bush, labor, affirmative, wage, enforce run ending for seed: Lerm no. 1 would be: chavez (score = 4) Lerm no. 3 would be: sweeney (score = 2,66) Lerm no. 3 would be: sweeney (score = 2,5)

Figure 3: Run-time log (part I).

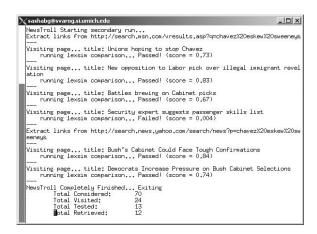


Figure 4: Run-time log (part II).



Figure 5: System progress.

Edit Vers Figurates Loois i			Links *
www.usatoday.com		0 0	
a www.nytimes.com	0 0	00 0	
www.msnbc.com	0 0	0	
dailynews yahoo.com		000	
		Time	of 13 selected
Available source(s)	time range Selected source(s)	Time from	1
dailynews yshoo.com	> >> <	Time to	
Highlight selected	Reset to default	Summarize selected Summarize all	

Figure 6: Cluster visualization.

Eck	Vew Pavortes	Tools Help	(1)				
	www.usatoday	com					
Source	www.nytimes	com					
Sol	www.msnbc	com					
	dailynews yahoo	com				10.010	
				т	ime		
						10 out of 1	5 selected
Av	your preferred sour adable source(s) ww.msnbc.com	ces and time	Selected source dailynews yahoo www.nytmes.com	com n	Time from		-
L		- ee	www.usaloday.co	HUL .	Time to		-0
	Highlight sele	cted	Reset to default	Summ	arize selected	Summarize all	1

Figure 7: Selected articles.

File Edit View Favorites Tools Help	
An An An An An An	
Summary @ 5% of the URLs that you selected:	
http://prem.s.tuench.edu/of.abs/10002550/http://www.second/s/g000004/dowra_words_11abs/170402 demo/data/010002550Abs/16/abs/www.second/s/g000004/dowra_words_relation demo/data/010002560http://data/second/s/g0000104/dowra_words_relation demo/data/01002550http://data/second	629 612 B-BU3H html 970942516 ABO html 970942509 EAV html970942507
http://www.scianci.edu/ciai/that/inserve-deno/data/0000224/http://www.sciance.com/poelios/sciai/ciai/ciai/ciai/ http://www.sciance.com/ciai/ciai/that/inserve-deno/data/00002254/http://www.usatoday.com/wwwis/odawa077 http://www.usatoday.com/wwwis/ciai/that/inserve-deno/data/00002254/http://www.usatoday.com/wwwis/oda200/bush	htm.978942586
http://jeruna.scianick.edu/clashindu/inasowa-daeo/dedu/clio002240.http://www.priae.com/productivationed/WTWEE http://jeruna.scianick.edu/clashindu/inasowa-daeo/dedu/00032543http://www.usadoige.org/norma/inaso/dedu/0012 http://jeruna.scianick.edu/clashindu/inasowa-deeo/data/010032543http://www.usadoige.org/norma/inaso/data/0100325 http://jeruna.scianick.edu/clashindu/inasowa-deeo/data/010032543http://www.usadoige.org/norma/inaso/data/01003 http://jeruna.scianick.edu/clashindu/inasowa-deeo/data/010032543http://www.usadoige.org/norma/inaso/data/01003	htm.978942586 142 htm.978942552
http://jenus.ai.uuich.edu/clushhu/inazowa-dono/dda.u010002240.http://www.optiona.com/populariv.kosiond/WTWFB http://penus.ai.uuich.edu/clushfurazowa-dono/dda.u010002240.http://www.uuadodg.entories/ddsau011 http://penus.ai.uuich.edu/clushfurazowa-dono/dda.u010002240.http://www.uuadodg.entories/ddsau01 http://penus.ai.uuich.edu/clushfurazowa-dono/dda.u010002240.http://www.uuadodg.entories/ddsau01 http://penus.ai.uuich.edu/clushfurazowa-dono/dda.u010002240.http://www.uuadodg.entories/ddsau01 http://penus.ai.uuich.edu/clushfurazowa-dono/dda.u010002240.http://www.uuadodg.entories/ddsau010000240.http: http://penus.ai.uuich.edu/clushfurazowa-dono/dda.u010002240.http://www.uuadodg.entories/ddsau01000040.http: http://penus.ai.uuich.edu/clushfurazowa-dono/dda.u010002240.http://www.uuadodg.entories/ddsau01000040.http: http://www.uuadodg.entories/ddsau010002240.http://www.uuadodg.entories/ddsau01000040.http: http://www.uuadodg.entories/ddsau010004040.http: http://www.uuadodg.entories/ddsau010004040.http: http://www.uuadodg.entories/ddsau010004040.http: http://www.uuadodg.entories/ddsau010004040.http: http://www.uuadodg.entories/ddsau010004040.http: http://www.uuadodg.entories/ddsau01004040.http: http://www.uuadodg.entories/ddsau0100404040040040040040040040040040040040	htm 978942586 142 htm 978942552 Score
http://jeruna.scianick.edu/clashindu/inasowa-daeo/dedu/clio002240.http://www.priae.com/productivationed/WTWEE http://jeruna.scianick.edu/clashindu/inasowa-daeo/dedu/00032543http://www.usadoige.org/norma/inaso/dedu/0012 http://jeruna.scianick.edu/clashindu/inasowa-deeo/data/010032543http://www.usadoige.org/norma/inaso/data/0100325 http://jeruna.scianick.edu/clashindu/inasowa-deeo/data/010032543http://www.usadoige.org/norma/inaso/data/01003 http://jeruna.scianick.edu/clashindu/inasowa-deeo/data/010032543http://www.usadoige.org/norma/inaso/data/01003	htm 978942586 142 htm 978942552 Score
http://jerum.a.sumit.e.4a/clashtat/inazores-dees/deta.010002240.http://www.orgines.com/populas/indusiond/WTWEE http://jerum.a.sumit.e.4a/clashtat/inazores-dees/deta.0100022540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.0100022540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.0100022540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.0100022540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.0100022540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.0100022540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.0100022540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.0100022540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.0100022540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.0100022540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.010002540.http://www.usadoge.org/normer/desau071 http://jerum.a.sumit.e.4a/clashtati/inazores-dees/data.010002540.http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://www.usadoge.org/normer/desau071 http://wwwwwwww	htm 978942586 142 htm 978942552 Score
http://jerum.a is unich.eks/dishthal/inasores-deso/deks.001003250.0htp://www.optimes.com/position/indiso/dfWTWE http://jerum.aisunich.eks/dishthal/inasores-deso/deks/0010032540.htp://www.usadoige.org/norma/disasor/f/ htp://jerum.aisunich.eks/dishthal/inanows-deso/deks/0010032540.htp://www.usadoige.org/norma/disasor/f/ htp://jerum.aisunich.eks/dishthal/inanows-deso/deks/0010032540.htp://www.usadoige.org/norma/disasor/f/ htp://jerum.aisunich.eks/dishthal/inanows-deso/deks/0010032540.htp://www.usadoige.org/norma/disasor/f/ htp://jerum.aisunich.eks/dishthal/inanows-deso/deks/0010032540.htp://www.usadoige.org/norma/disasor/f/ htp://jerum.aisunich.eks/dishthal/inanows-deso/deks/0010032540.htm Seleves.tbs-interess-in.sensia/huwasmeel.terevals in making the 'United States' = nation of crybabies' and that ''with m disagot of orch-which husti aft no surgene that sfew stab. tock.	htm 978942386 142 htm 978942352 ten are so often 7.39 0.00
http://jeron.a.sunit.e.tai/itainthu/inasores-deso/detx.010002500.http://www.orgines.com/position/itainteed/WWWE http://yeron.a.sunit.e.tai/itainthu/inasores-deso/detx.010002540.http://www.usadog.com/itaintee/it	htm 978942386 142 htm 978942352 ten are so often 7.39 0.00
http://jerum.a.sumich.edu/clashhal/inazores-dees/d4a.010002240.http://www.opfase.com/populasin/action4077WTBP http://jerum.a.sumich.edu/clashhal/inazores-dees/d4a.010002240.http://www.usadog.com/serves/d5asu077 http://jerum.a.sumich.edu/clashhal/inazores-dees/d4a.0100022453.http://www.usadog.com/serves/d5asu077 http://jerum.a.sumich.edu/clashhal/inazores-dees/d4a.0100022453.http://www.usados.com/serves/d5asu077 http://jerum.a.sumich.edu/clashhal/inazores-dees/d4a.0100022453.http://www.usados.com/serves/d5asu077 http://jerum.a.sumich.edu/clashhal/inazores-dees/d4a.0100022453.http://www.usados.com/serves/d5asu077 http://serves/doi.on/serves/dees/d4a.010002453.http://www.usado.d0/yerves/vois2000.btush http://serves/doi.on/serves/dees/d4a.010002453.http://www.usado.d0/yerves/vois2000.btush http://serves/doi.on/serves/dees/d4a.010002453.http://www.usado.d0/yerves/vois2000.btush http://serves/doi.on/serves/dees/d4a.010002453.http://www.usado.d0/yerves/vois2000.btush http://serves/da.d0/serves/da.d0/yerves/vois2000.btush http://serves/doi.on/serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/vois2000.btush http://serves/da.d0/yerves/da.d	htm 970942586 142 htm 970942552 seen are so ciffen 7 39 600 iseen 0.00 0.00
http://jerum.a.sumich.edu/dashhal/inazores-dees/dedu/di002240.http://www.orgines.com/production.doc/07/WTBP http://jerum.a.sumich.edu/dashhal/inazores-dees/dedu/00023420.http://www.usadog.com/twws/voles/0007042 http://jerum.a.sumich.edu/dashhal/inazores-dees/dedu/00023420.http://www.usadog.com/twws/voles/000704ah hummarizing.ddf sectores.gl_3% = 32 sectores hummarizing.ddf sectores.gl_3% = 32 sectores faileres the increase in securd harsement formals in stating the United States " s nation of crythebier" and that " with m the larger of such witch hunds if is no surgices that a few stales buck. gov Freelest.Back Buck Transition - includes news staties and analysis profiles of Column increases and Wata House adm Freelest.Back Buck Transition - includes news staties and analysis profiles of Column increases and Wata House adm Free Back Douland Harvellow	htm 970942586 142 htm 970942552 seen are so offen 7 39 0.00 isees 0.00 0.00

Figure 8: Summarization interface.