Tracking Information Flow in Finance Text

Author: Will Radford wradford@it.usyd.edu.au
Supervisor: James R. Curran and Ben Hachey (Macquarie University and Capital Markets CRC)
School of Information Technologies

1. Introduction

Financial news plays a central role in how people interact with the market.

- Companies must continuously disclose any "price sensitive" information to the exchange, which published a primary source of official announcements.
- News agencies report on these (and other) events and journalists may add value by presenting background knowledge, expert analysis or editorial commentary.
- Participants who can absorb and react more quickly to information are at a substantial advantage in the marketplace.
- Tracking information flow means that information can be more effectively linked, filtered and presented.

2. Problem

We study Australian Securities Exchange (ASX) official announcements and Reuters NewsScope Archive (RNA) stories covering the same companies and time periods.

Given an announcement and a story ASX-RNA pair, how well can we identify the following:

REL - information from ASX is in RNA

FACT - the information is *new*

BACK - RNA contributes background information

ANLY - RNA contributes analysis or commentary

3. Approach

We frame Information Flow as a supervised text categorisation problem. Specifically, for each ASX-RNA pair, which (if any) of REL, FACT, BACK or ANLY are true. We had finance students annotate (see Figure 1) 39,869 ASX-RNA pairs to create the data which we use to train and test combinations of features, including textual similarity approaches from:

- Plagiarism Detection [6]
- Information Retrieval [4]
- Text Reuse [3]
- Topic Detection and Tracking [1]
- Novelty Detection [5]

4. Modelling Information Flow

We define and extract the following features.

∩ **ngrams** unigrams and bigrams that appear in both ASX and RNA

ngram differences unigrams and bigrams which are in ASX and *not* in RNA (and vice versa)

similarity scores over text, title and number tokens

sentences number of whole sentences that appear in both ASX and RNA

sequences number of common subsequences in both ASX and RNA

precision indication of the matching numbers of different precisions: e.g., $500 \rightarrow 000, 123 \rightarrow ###$

time lag elapsed time between ASX and RNA represented as bins of increasing size (can be negative)

time of day ASX and RNA timestamp to nearest 30 mins

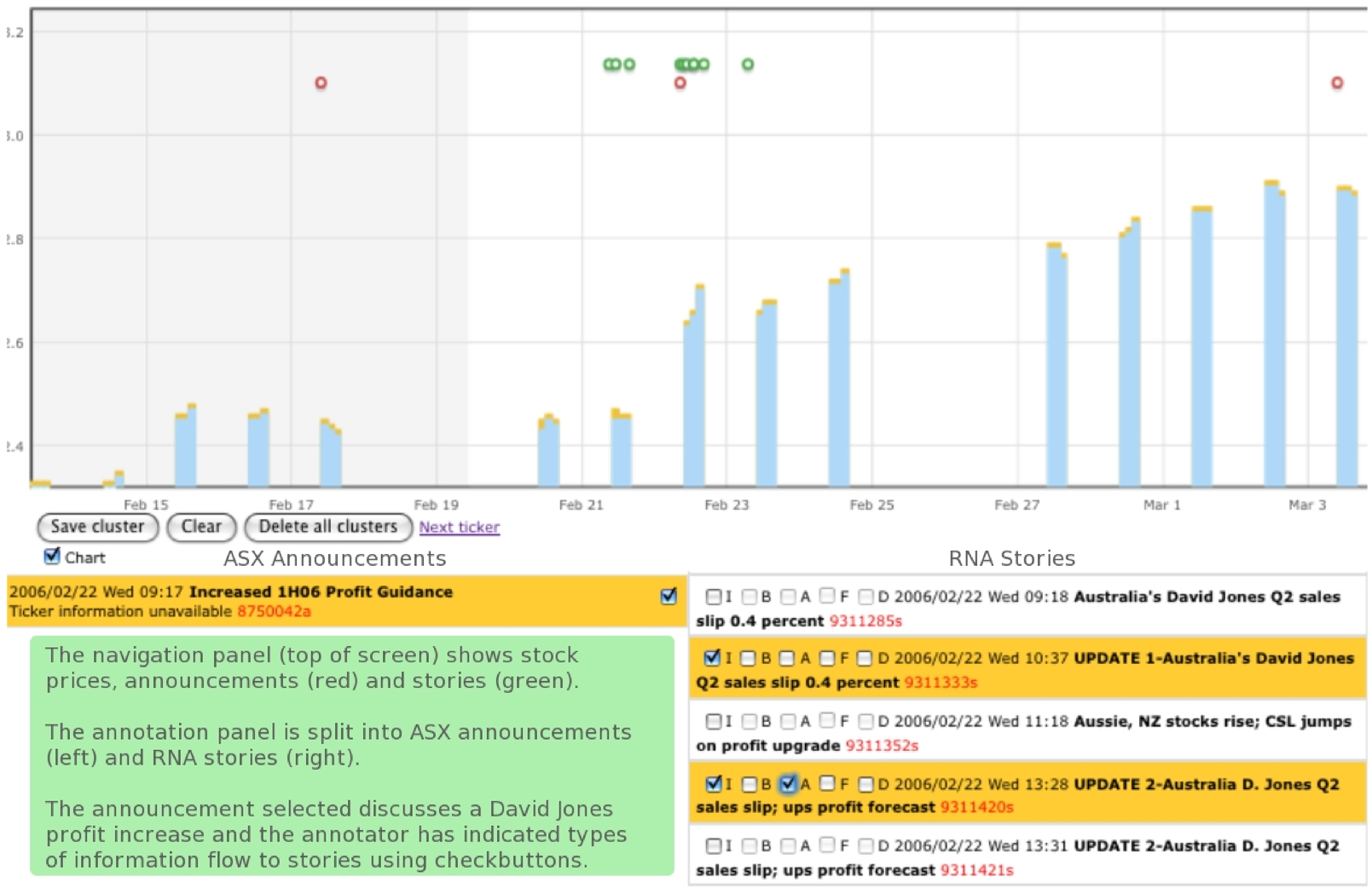

Figure 1: *Screenshot showing* ASX-RNA *information flow*

5. Results

We run ten-fold cross validation experiments for each of REL, FACT, BACK, ANLY using the `megam` [2] Maximum Entropy classifier. Our baseline is the bag-of-words taken from the intersection of the ASX-RNA title and text.

Link	Baseline	System
REL	78.6	**89.5**
FACT	52.7	**73.4**
BACK	74.4	**85.7**
ANLY	65.5	**80.4**

Figure 2: *F-score for classification experiments*

Subtractive analysis is used to assess whether separate features had a statistically significant impact.

Link	REL	FACT	BACK	ANLY
∩ text-1g	·	·	·	·
∩ text-2g	⋆⋆	·	⋆⋆	·
∩ title-1g	·	·	·	·
∩ title-2g	·	·	·	·
ASX \ text-1g				
ASX \ text-2g				
ASX \ title-1g	⋆⋆		⋆⋆	
ASX \ title-2g	⋆⋆		⋆⋆	
RNA \ text-1g		·		·
RNA \ text-2g		·		⋆⋆
RNA \ title-1g	·	·	⋆⋆	·
RNA \ title-2g	⋆⋆	·	⋆	·
Text sim^y	⋆⋆	⋆⋆	⋆	·
Title sim^y	·	⋆⋆	·	·
Num sim^y		·		·
Sentences	·	·	·	·
Sequences	·	⋆⋆	·	·
Number $prec^n$		⋆		·
Time lag	⋆⋆	⋆⋆	⋆⋆	·
Time of day	·	·	·	⋆

Figure 3: *Feature combinations for the best performing experiments.*
· marks active, stars mark significance at ⋆ ($p<0.05$) and ⋆⋆ ($p<0.01$)

6. Future Work

- Examine more sources – e.g., blogs, forums
- Investigate logical inference – can deeper meaning be extracted from the stories?
- How does a story unfold, within RNA and also the wider news ecosystem?

7. Conclusion

We demonstrate that we can feasibly track information flow in finance text and make the following contributions:

- Annotation scheme that codifies information flow and three types of journalistic contribution that can be applied with high agreement.
- Exploration of features from diverse fields to model information flow.
- Classification of information flow at 89.5% F-score and journalistic contribution from 73.4% to 85.7% F-score.
- Favourable agreement with *majority* decision of 86% F-score, our system scores 77% for flow classification

8. Acknowledgements

This research was conducted in conjunction with the Capital Markets CRC under a High Achiever's Scholarship.

References

[1] James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45, New York, NY, USA, 1998. ACM.

[2] Hal Daumé III. Notes on cg and lm-bfgs optimization of logistic regression. Aug 2004.

[3] Donald Metzler, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat, and Justin Zobel. Similarity measures for tracking information flow. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 517–524, New York, NY, USA, 2005. ACM.

[4] Gerard Salton, A Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[5] Ian Soboroff. Overview of the trec 2004 novelty track. In *The Thirteenth Text Retrieval Conference (TREC 2004)*, 2004.

[6] Michael J Wise. Yap3: improved detection of similarities in computer program and other texts. *SIGCSE Bull.*, 28(1):130–134, 1996.

Capital Markets CRC Limited

The University of Sydney