Searching Real-Time Financial News: PAR for the Course
by Lawrence C. Rafsky

Real-time news (also known as live news, streaming news or breaking news) – especially news focused on business and financial matters – is widely read and builds up quickly. A typical newspaper/newswire aggregation system will process three stories per second around the clock, adding approximately 250,000 stories every business day to the collection. Large commercial systems substantially surpass this total.

The needs and behavior of end-users searching collections of this type differ from general searching norms in several key aspects:
- The computational burden is closer to the classic alert/routing problem than it is to the ad-hoc search problem.
- Business work is about the opportunities of today and tomorrow.
- All news is structured.
- User queries do not typically consist of a few terms.
- Users have a high degree of topic familiarity and topic focus.

The Computational Burden

The computational burden is closer to the classic alert/routing problem than it is to the ad-hoc search problem (these are standard terms-of-art in IR – as, for example, Robertson [1] explains). In fact it closely matches the batch routing problem, because (1) users save searches and re-execute them frequently (rather than entering searches in an ad-hoc manner) and (2) very little latency is tolerated by users. A news article arriving in the last few moments should be returned as a hit if the article is a search match to a just-executed query.

Combine this low-latency requirement with a saved-search, re-execution cycle that is often a minute or less (programmed as web page auto-refresh), and the similarity with alerting becomes obvious. Thus the old folk-theorem, “In search you have all the time you need to study the archive, and no time to study the query; in alerting you have all the time you need to study the query, and no time to study the archive,” doesn’t hold here.

However, real-time news is still search. A static set of matches (in this case, news headlines and summaries) is returned, and the user demands some ranking of the results by meaningfulness and pertinence. In the alert problem it is sufficient to present the user with a temporal stream of matches; not so here.

Thus real-time news search demands the computational agility of alerting and the semantic processing of search. The job is made somewhat easier by the batch nature of the query collection – the saved searches. In the commercial system NewsEdge™, of the last million searches (looking backward from August 15, 2009), approximately 96.5% of executed searches were repetitive submissions of stored queries. This statistic excludes users who specifically posted alerts and received a content push from NewsEdge – it refers only to users who requested (perhaps automatically) and received a web page of results.

The Opportunities of Today and Tomorrow

For the most part, real business work (as opposed to academic business study) is about the opportunities of today and tomorrow. Searchers care less and less about stories as they come from earlier and earlier time points. Thus “article publication date” is not merely a time facet or a temporal query profile [2]; it is a fundamental part of the search process.

Furthermore, there is a fundamental time flow to news: themes develop,
expand and mature over time. Arriving news stories reflect this rhythm. Thus news temporal order has a fundamental effect on how a retrieval system handles clustering, novelty discovery and redundancy [3].

One common approach is to enforce a time cut-off, to limit retrieval to, say, the last 180 days (unless specifically directed otherwise by the user) and then proceed normally with all calculations such as relevance. Another approach is to use story time (more correctly, age) directly in relevance formulas, down-weighting older stories.

Taking a cue from the similarity between the low-latency news search problem and the alert problem, some systems display results only in reverse chronological, or time, order, and attach a relevance value to each headline. A twist on this, which is perhaps more user-friendly, is to flag or star stories – still displayed in time order – that have a relevance score above a threshold.

The threshold can also be used as a filter, keeping the story hits in time order but only showing those that have been relevance-flagged.

All News Is Structured

All news is structured – whether or not the document arrives in html, xml or plain text. News – by definition – has at least a headline and one or more of the following: deckline (sub-headline), byline (author and location) and body (the story itself). And like all documents there is a source and potentially a copyright, but these features are not distinguishing characteristics of news. Moreover, news articles from printed newspapers – and some websites – have typeface weight and page/section placement data that convey valuable information, which has been exploited in a number of IR systems, although we are covering unified collections of news here, including newswire and blogs that lack this structure. There can also be summaries or leads (similar to abstracts in scholarly papers) – or, if not, these items can be synthesized by exploiting the top-down structure of news writing (more on that below) and metadata sections containing such information as company identifiers.

This structure can and must be exploited in relevance and novelty/redundancy calculations. Most work on IR systems for structured documents focuses on segment (as opposed to entire document) retrieval, and techniques such as those for computing relevance for xml documents or for defining term weights within segments are very valuable in real-time news retrieval. An article by Weigel, Schulz, & Meuss [4] is one of many possible references on these topics.

Beyond simple structure, sophisticated relevance calculations should take into account that editorial rules and conventions (writing styles) exist, are tightly enforced for some of these structures, and they should therefore weight terms accordingly. Reading Ellis’ Copy-Editing and Headline Handbook [5] is eye-opening for anyone writing a news retrieval system.

But the world of journalism is not a simple one. New websites are setting new rules, although the story concept remains the same. The top-down nature of news stories also comes into play when computing relevance, except for “Summary,” “Headlines of the Hour” or NIB (news in brief) stories, since they consist of multiple small stories knitted together (the print media equivalent of an RSS file, except often without links to longer stories).
User Queries

User queries in the real-time news environment do not typically consist of a few terms. The “2.4-terms-per-query” rule [6] simply does not hold. Taking a random sample of (roughly) 50,000 saved searches on the NewsEdge system, we find an average of 44 terms per query. This characteristic has profound implications for relevance calculations that use term frequencies (as nearly all do, in some capacity or another). Inspection shows that many of these terms are proper nouns, in particular, company names (or company ticker symbol identifiers), whose presence make perfect sense, since a list of company names is a proxy for an industry or directly represents an investment portfolio or sales territory. In the case of an industry query, it might be argued that a metadata value (an industry category) would be a better (single term) query, but not everyone defines industries the same way, and end users very much prefer the definitive nature of an explicit company list (usually with additional keywords and phrases).

Unlike most OR clauses that a search algorithm would encounter, here the expected number of term matches is often one (of, say, 44). That more terms may match in no way implies greater relevance. In fact too many matches would indicate some type of boring list or table of companies, the kind of story nobody wants to read.

Topic Familiarity

Users in our environment have a high degree of topic familiarity [7] and topic focus. This characteristic dramatically affects the use of relevance criteria. No one in the business world (we are excluding academic business research here) searches business news for fun – only for profit. Give the user high precision but low recall, and he misses opportunities, losing money. Give the user high recall but low precision, and you waste his time. Luckily the topic focus, combined with the structure analysis discussed above, helps with both recall and precision.

Consider a story on a concert to be held at Verizon Center in Washington, D.C., with a box office that accepts Visa, MasterCard and American Express. It is unlikely that a business searcher wants this story returned as a highly ranked hit for a query that included the company symbols NYSE:MA, NYSE:VZ and others. But it has two perfect term matches. Possible saving graces include:

1. It is not a business story.
2. The phrase “Verizon” appears next to “Center,” and that phrase only appears once (so it is most likely not a story on Verizon’s investment in the naming rights, which came along with the acquisition of MCI Worldcom).
3. The term “MasterCard” appears at the bottom of the story.
4. The ticker symbols (NYSE:MA and NYSE:VZ) do not appear in the metadata record that accompany the story.

There are other cautionary notes for search engineers, however. We need to worry about a story that has Verizon offering cell phone deals to MasterCard cardholders – highly relevant – and snippets like “work at Verizon centers on increasing renewal contracts,” which reveal problems with ignoring case and conflating singular and plural versions of terms.

Some IR systems are designed to maximize diversity of results; that is, when a search term is ambiguous they try to return stories covering all possible meanings. The topic focus of business news retrieval systems eliminates this requirement. We can be assured that the user with a query containing “poultry stocks” wants hits on Perdue, not cooking. Recipes would routinely be included in stories retrieved for commodities traders were it not for the recognition of topic focus.

Alternatives for Defining and Calculating Relevance in the Business News Environment

It should be clear from the above discussion that standard, classic, vector-space relevance, which is calculated by the formula “term frequency multiplied by inverse document frequency” (tf*idf), will not be sufficient in and of itself for business news retrieval systems. Even when we take document structure, query length and topic focus into account, more is needed to do a good job for working business news searchers.

First, users this sophisticated understand Boolean logic and will expect all hits to be matches. This expectation is not really at odds with relevance-
based retrieval, since even Salton suggests that Boolean and relevance search can be combined by returning all true Boolean matches, computing the relevance score for each match and then sorting (or flagging so that stories can be kept in reverse chronological order, as discussed above) [8, Chapter 10]. This procedure is routinely carried out in nearly all business news retrieval systems in the commercial marketplace.

But we know now that other measures – not based on closeness to the query – need to be brought to bear. Three main alternatives to classic relevance have surfaced (in addition to using print and web position and placement clues, as discussed above, which we will not pursue further here):

a) Evaluate novelty of results; promote matches (flag or sort to top) that bring new information to light. [9]

b) Derive ranks based on authority, using either author analysis [10] or link analysis of the collection. (For example, see [11] or [12].) Of course this approach includes Google-like inbound link ranking, usually attributed to the founders of Google but priority would appear to belong to Yahong (“Robin”) Li, the founder of Baidu, alone – based on Li’s paper at the 1997 Philadelphia SIGIR workshop [13], the date of Li’s patent filing (1997, 5 February) [14], and the submission date of Li’s IEEE Internet Computing paper [15].

A weaker version of authority is popularity, used on social media sites, measuring “clicks” (in this context, reading the full story after being presented only with the headline and summary – “most read” voting) or simply counting how many readers comment on a story.

c) Attempt to figure out the user “use case” – the context of the search. What is the user going to do with the results? How can the search results help with the “work task?” [16] Elevate those matches that help the most.

We can dismiss (b) almost immediately; it is of no help in this problem since real-time news contains neither the links nor the author attributes necessary to fuel this ranking engine. And if we consider popularity measures, there are two problems. The first is due to the low-latency requirements since ranks need to be computed in real-time. The system simply cannot wait and collect “story read” information from users before ranking the matching stories that just arrived. Second, while such information and more could be collected and used to inform the rankings in subsequent searches, user-behavior privacy issues – vital in a competitive business setting – make any attempt to collect (much less mine) such data ill-advised.

Approach (a) is very likely to fail in this context as well. Novelty analysis depends vitally on thematic isolation. The thinking behind much novelty analysis is that if the theme is new the match provides novel information. This assumption fails badly in financial news, however, since financial events unfold over time in unexpected ways. Here are four (hypothetical) financial news threads – the headlines in each thread are presented in time order, and the key stories (the ones that deserve high rank) are shown in bold.

- Company XYZ announces its earnings results will be available on July xx
- Company ABC has begun a clinical trial on Wonder Drug ZZZ
- Company ABC announces successful results from its ZZZ clinical trial
- Company ABC wins FDA approval for ZZZ
- Company P and Company Q are rumored to be in merger talks
- Company P and Company Q to merge
- Company P and Company Q announce a merger
- Company P and Company Q complete their merger
- Company Z will have to file for bankruptcy protection if things don’t improve
- Company Z makes an initial bankruptcy filing
- Company Z bankruptcy hearings proceed

Some interpretation of novelty should indeed be a key component of search-result ranking for real-time business and financial news, but “new theme” is not the answer.

Eliminating the options above leaves us with (c) – rank the stories on how we believe the stories might or will influence the user’s actions – the eventual actions the user takes (or decides not to take) after studying the
search results presented. Van Rijsbergen has put it beautifully – this (abridged by us) quote is worth studying:

The nature of what is wanted by a user is a matter for debate. . . . It is not enough to say that what is wanted is a matching item, matching items may be irrelevant or useless. . . . It is a convenience to conflate relevance with aboutness, but now especially in the context of Web searching it may be necessary to begin separating these. . . . [I]ncreasingly searches are done within a context of performing a task; the nature of the task could have significant effect on what is worth retrieving [6, p. 13].

For example, I need to find the exact address of a hotel in a city foreign to me. Typing the hotel name into a web search engine is usually of little value – it has been assumed (either by the wisdom of crowds or some hidden commercial interest) that I want to make a reservation at the hotel, and often I have to wade through pages of discount hotel booking site URLs to find the hotel’s actual website. This mistake is one of “use.”

And this is our suggestion: Presume a particular action response on the part of our business/finance news searcher. We call this technique PAR searching – Presumed Action Response.

The key observation here is that the world that invests in business is a parallel, shadow world to the one that operates and cares for business. A person in business searching for news on Company Big Widgets might care about the following:

- competing with Big Widgets
- selling to Big Widgets
- buying from Big Widgets
- investigating the executives who work for Big Widgets
- getting a job with Big Widgets
- suing Big Widgets
- determining how Big Widgets is doing financially (since that affects all the above)

or something similar. A person investing in business searching for Company Big Widgets cares about the exact same things, since they expose the levers that determine whether investments in Big Widgets will make or lose money.

For publicly traded companies the stock market provides insight retrospectively as to whether a news story on a particular company “moved the market” for that company’s stocks or bonds – which means the news story in question should have been highly ranked if the searcher had issued a search on that company name just as the story was received by the system.

Therefore, using historical data analysis, we should be able to determine those features in a story (on a public company) that cause it to “move the market.” We are not suggesting this analysis is an easy undertaking – there are a number of issues and problems with such studies, touched on below. But we do claim that it is “Job One” for developing a search-result ranking algorithm for real-time business and financial news.

Having found these market-moving, news story features – words and phrases – in public company stories, we can extend the analysis *ipso facto* to stories on non-public companies and more generally to non-company-centric business stories, since these stories will contain similar features (using standard IR similarity formulas but paying attention to the structure issues discussed above).

We define the impact of a story as some quantitative scoring that measures the presence or absence of such features. The exact numerical scales used and other details are not important for this paper – what is important is this: In the domain of real-time business and financial news, we believe impact should form the basis of result ranking.

Observe that classic tf*idf calculations still play a key role, but indirectly – they are used in similarity measurements for stories that cannot be related to movement in public markets.

(Note: Government economic-condition news releases move interest rate and foreign exchange markets in ways similar to the movements produced by breaking stories on public companies. Our methods thus extend to them as well, although we do not concern ourselves with this type of story further in this paper.)

The concept of “moving the market” is not trivial. It does not necessarily mean a stock price has changed. Often one finds that, in reaction to a major story, there is wild trading, but the price moves very little on large volume. In these cases, volatility is the right measure (and we look for changes in
volatility before and after the issuance of a news story). See Alexander’s book on market models [17], Chapters 3 and 11. To assess impact we carefully measure over time the actual changes in markets when stories break, using time-series statistics, and isolate the features involved. Here are some of the factors we weigh for news on public companies:

- Is the timing of the news event known in advance (earnings reports), or is the fact that there is news itself a surprise, representing “true novelty” (patent infringement case, for instance)?
- How close are we to the next earnings announcement date?
- How close are we to the next options expiration date?
- Has volatility decreased across the time of the news event since uncertainty has been reduced, or the opposite?
- Has volume dried up ahead of the news release since uncertainty has increased the real (unobserved) bid-ask spread to the point where there are no trades to be done?

Having in this way assessed (actual) impact for a large number of news stories on public companies, we then run a standard categorizer (a separating hyper-plane algorithm or a support vector machine), eventually building a linguistic feature set, a set of linear rules and a threshold computation that can be applied in real-time to any story (whether on a public company or not). Distance (on either side) from the threshold can be turned into a numeric measure taking values from 0 to 1 of (predictive) impact and be used to rank stories.

The method appears to work well. Here is a screen shot from our test bench.

Predicted impact is shown as a number from 1 (low) to 9 (high) – just a re-scale and rounding of the actual continuous value. The query was: Biotech Earnings. Note that this is not a query on a public company name.

You can see that new earnings reports got high impact scores (mostly 9s, one 7), a “reiterate” research report got a 6, a research report with little to say got a 2 and (most impressively) an announcement of an earnings conference call – an “announcement about an announcement” – was marked down to an impact ranking of 1.

As long as our assumption on end-user actions is correct, we believe this PAR approach, using our impact measure, produces meaningful rankings of Boolean-match results for working users of real-time business and financial news search systems.

PAR can be viewed another way. Since investors do actually read news stories and trade securities on the basis of what they read, if we have accurately captured their actions (in the sense that we can determine that a market movement resulted from the issuance of a news story), then impact is in fact relevance feedback from an informed community. It is not direct feedback, but a proxy. Hence the other meaning of PAR: Proxy Activated Relevance-feedback. It is usually easy to focus on a graphic rather than a number, so in our commercial NewsEdge products we are introducing impact as a bar graph:

Note the IBM story with no impact. We are going to allow both a sort on impact and a filter (show no stories below an impact threshold).

We believe we have done our job. We have made PAR.

Acknowledgements

The author wishes to thank the members of the NewsEdge team at Acquire Media, and in particular Kristen Carney, Zhi Chen, Boris Guralnik, Dan O’Connor and Nancy Yee (alphabetical order). Thanks are due as well to Irene Travis, editor of the Bulletin, for suggestions on the first draft that substantially improved the exposition.
Resources Mentioned in the Article


