Map-Based Newsfeed Visualisation in Support of Sensemaking

Literature and Technology Survey

Damask Talary-Brown

Bachelor of Science in Computer Science with Honours
The University of Bath
2016

Contents

Literature and Technology Survey

1.1 Why don't we understand the news?

The day the result of the 2016 United Kingdom EU Membership Referendum was announced, the Google Trends Twitter account reported a 250% increase in searches for "What happens if we leave the EU?" [23]. Much like the case of David Leonhardt's article for the New York Times in 2008 which began, "Raise your hand if you don't quite understand this whole financial crisis," national news commentary had focused on little else in the preceding months. It has become apparent that prolific coverage alone is not enough to engage and support the public in understanding current events.

Historically, news media has been limited in the volume of content it can produce by physical constraints such as printing costs, but the rise of the internet as a platform to deliver it has lead to an explosion of content, both through existing media channels and through competing social media websites and blogs. The term *ambient news* was coined by Hargreaves and Thomas [7] to describe the ubiquity of news in the current information landscape. Others have commented in a more critical light; describing the proliferation of competing news media as "as pervasive–and in some ways as invasive–as advertising." [13, p.2]

In 2007, The Associated Press conducted an extensive field study [1] into the news consumption habits of young adults. Among their key findings were three points which essentially summarise the news overload issue;

- **"Consumers are experiencing news fatigue."**

  The study found participants were debilitated, and that their levels of dissatisfaction lead to a decrease in the effort they put into news acquisition. This is consistent with multiple other studies [10, 17, 5] which found participants across every demographic were overwhelmed by the amount of news content available to them and agreed it prevented them exploring news on less familiar topics.

- **"Story resolution is key."**

  Participants' consistent enjoyment of sports and entertainment news was due in part to the formulaic storytelling which characterises these types of journalism, with clear chronology to provide contextual back story. The feeling of enjoyment gained from reading procedural stories directly contrasts with what the same participants experienced reading World news, where they struggled to find resolution to stories which were unfolding at the time.

- **"Consumers want depth but aren't getting it"**

  It was observed that participants, in their efforts to discover *below-the-fold* content (defined in the context of the AP's model [1, p.37]) from particular headlines, often found themselves reading the same summary-level content from different news sources. A recommendation was made that news providers should be"designing innovative formats and creating easier pathways to deep content" [1, p.49] in order to support this.

Initially, the third point seems to be a direct contradiction to the first; we are overwhelmed by the volume of news we are exposed to, but we also crave more detail from the news we do consume. However, it brings to light the issue of information *quality* as a requirement of news consumers.

Journalism, and therefore its quality, can be viewed along a spectrum between two models; a model for the communication of facts, and a model for entertainment and storytelling. From the three points above, it is clear that quality at both ends of the spectrum is being sought, since the desire for quality below-the-fold content is covered by the first model, and the desire for quality story resolution by the second.

1.1.1 Information Overload

News fatigue is a domain-specific type of information overload, a phenomenon formally defined as "when the information processing demands on time to perform interactions and internal calculations exceed the supply or capacity of time available for such processing." [20, p.206] Information overload is a multifaceted problem which can be modelled as a combination of three contributing factors (Figure 1.1).

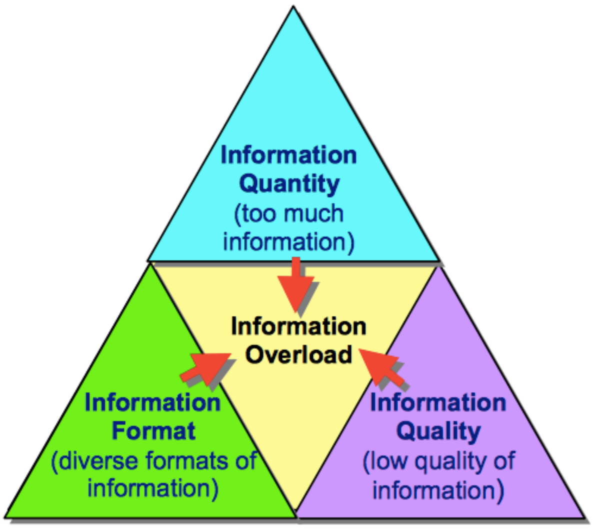
Information Quantity (too much information)
Information Overload
Information Format (diverse formats of information)
Information Quality (low quality of information)

Figure 1.1: Dimensions of information overload, as defined by Ho and Tang [9].

These factors correspond directly to the three points previously identified from the Associated Press study [1]. Information quantity causes news fatigue, information format determines the level of possible story resolution, and information quality determines how much depth a reader can gain from the news they consume. The authors did not find a single solution which could address all three factors, but they did identify information quantity as the most significant contributor to overload.

Bergamaschi et al.[3] further decomposed information quantity into spatial and temporal dimensions in the specific context of news articles. Spatial quantity refers to articles which are near-identical in terms of facts presented being published by different media outlets, and temporal quantity refers to articles on a single topic being published and updated in quick succession over a short period of time.

To adequately address news fatigue, all four dimensions of information overload should be considered, and will therefore be explored in more detail.

Information Quality

In the context of factual data rather than news specifically, Strong et al [22] defined information quality in terms of four components; intrinsic quality, accessibility quality, contextual quality and representational quality. If news can be rationalised to its core function as an interpretation of facts and other raw data such as images, then this same framework can be experimentally applied to news journalism in order to determine which factors could influence its quality.

Intrinsic quality is a measure of the accuracy, objectivity, believability and reputation of data. In the context of news, the first three factors would typically be true for all major news sources, and the reputation would be dependent on whether the article originated from a trusted source or not. Accessibility quality is less relevant to online news media, as it is concerned with data access and security. Contextual quality is the most relevant category in respect to news, concerning timeliness, amount of data, and value-added. In the news domain, this would mean an article's quality is dependent on its performance against a background of other articles; whether or not it contributes anything recent or previously unknown. Finally, representational quality is concerned with ease of understanding and interpretability, which are easily translatable concepts.

To summarise, applying the Information Quality Framework[22] to news articles suggest quality can be influenced by the reputation of the source, the timeliness of publication, value-added by the article (i.e. content which couldn't be derived from other sources) and ease of understanding.

Information Format

The domain of news articles is a more specific information space than that of documents in general, and by nature most news articles share some common formatting and structural elements such as headlines, timestamps, and relevant images. As a result of this, it is unlikely that any two articles from popular news providers would be diverse enough in content format to overshadow the information quantity problem.

Temporal Quantity

The rise of social media sites such as Twitter delivering news to consumers has lead to a high degree of news fragmentation, due to the constraints of the microblogging service's

140-character limit. 24-hour television news paved the way for new formats of real-time content delivery, and the ever-expanding network of online social media channels stepped up to deliver. It logically follows from the fragmented nature of real-time news journalism that temporal quality suffers; stories are published and updated intermittently over short periods of time, meaning there is more for the consumer to piece together to understand a story.

The fragmentation is somewhat mitigated by Twitter's use of hashtags to denote a Tweet's topics. Hashtags help readers form a coherent and picture of unfolding events from the incremental contributions of thousands of participating users. [4]

Murata and Phuvipadawat [16] developed a methodology to collect and group Tweets on breaking news topics, using hashtags for topic identification or *story-finding*, and grouping similar messages together to form a single news story. Their algorithm for similarity is a function of the TF-IDF [21] of the two messages and the number of named entities they have in common.

Spatial Quantity

It is in the nature of news that newsworthy stories get repeated across multiple sources. When consumers read news on a particular from more than once source, it is likely that they will read variations on the same facts in multiple articles.

Attempts such as [2] have been made to synthesise summaries of collections of similar online documents, a practice here termed *information fusion*, with news articles from different sources being given as a specific use-case. However, the process of extracting common sentences between documents was in order to reformulate them into a single summary, rather than to determine the level of similarity between the documents.

A more relevant approach was presented in [15], where the *title* and *description* attributes of elements in RSS feeds were used as content descriptors to mitigate the overhead of processing entire documents for phrases. This decision can be supported by other work such as [12], which found up to 50% of a news document's keywords can be found in its title. The content descriptors are then used to compute phrase n-grams as a measure of similarity between any two documents. The similarities in this case were used to remove subsumed articles and cluster non-redundant similar ones, in order to streamline feed content for readers.

It should be noted that there is an overlap between the notion of spacial quantity and one of the influencing factors in information quality; contextual quality. If a feed contains two articles which state the same number of identical facts, they therefore contribute to information overload on both the qualitative and quantitative fronts.

1.1.2 Supporting Sensemaking

Sensemaking is the basis for forming contextual knowledge; the process by which we incorporate new information into our existing cognitive frameworks, and how we go from reading something to understanding it. [14] In broader terms, Weick et al. describe sensemaking as "[being] about the question: What does an event mean? In the context of everyday life, when people confront something unintelligible and ask, 'Whats the story here?'" [24, p.85] This relates directly to the news overload problem because sensemaking describes the contextual story resolution that the Associated Press study [1] identified news consumers are craving.

Offering further insight into Leonhard's aforementioned New York Times article during the financial crisis, Journalism Professor Jay Rosen wrote in a blog post some months after the article was published, "there are certain very important stories – and the mortgage crisis is a good example – where until I grasp the whole I am unable to make sense of any part."[18]

It has also been observed that often readers are not interested in specific articles on a subject, and only the thematic content of the topic they belong to.[11] How then, do readers make sense of a collection of articles surrounding a particular topic?

When presented with a large document collection on any kind, people typically start by clustering the contents into groups which form a heuristic representation or mental model which can be used to provide an overview. [19]

Writing for the Columbia Journalism review in 2008, Nordenson outlined a suggestion for the new roles of journalism in the information era; "By linking stories to one another and to background information and analysis, news organizations help news consumers find their way through a flood of information that without such mediation could be overwhelming and nearly meaningless."[13, p.10]

Similarly, [14] makes the recommendation in the context of contemporary media consumption that news providers should adapt to an environment of news overload by adding facilities enabling readers to categorise, sort and search news collections. Additional findings of this study suggested that the contextual background provided by having more detailed coverage aids the sensemaking process, as it helps users form links between new information and their existing frameworks, but this presents an interesting conflict with the goal of reducing information overload.

A recognised technique for bridging the gap between a set of data and a user's mental model and subsequent of the data is information visualisation, or InfoVis. [25, 8].

Yi et al. [25] identified four overlapping InfoVis processes which describe how insight can be gained after sensemaking; *Provide Overview*, *Adjust*, *Detect Pattern*, and *Match Mental Model*. The Provide Overview process allows a reader to recognise what they know and what they don't know from the information they are processing. Adjusting allows them to change the level of abstraction or field of selection of that information. The Detect Pattern procedure is where structure and trends are found, whether expected or otherwise. Match Mental Model is where the links are formed between the new data and the users' existing cognitive frameworks.

Even though the structure and purpose news documents does not naturally suggest an appropriate visualisation for them, it is still the case that visualisation of non-spatial data can provide vital insight, especially when the quantity of data is vast. "Because we live and perceive in a physical world, it is easier to convey the information to the observer if the information is represented by being mapped to the familiar physical space."[6, p.39]

1.2 Designing Visualisations

1.2.1 Keyword Extraction

1.2.2 Topic Coverage

1.2.3 Maps

1.3 Evaluating Usefulness

1.4 Conclusions

Bibliography

[1] Associated Press and Context-Based Research Group. A new model for news: Studying the deep structure of young adult news consumption. `http://manuscritdepot.com/edition/documents-pdf/newmodel.pdf`, 2008. Accessed: 26/10/2016.

[2] R. Barzilay, K. R. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 550–557. Association for Computational Linguistics, 1999.

[3] S. Bergamaschi, F. Guerra, and B. Leiba. Guest editors' introduction: information overload. *IEEE Internet Computing*, 14(6):10–13, 2010.

[4] A. Bruns, T. Highfield, and R. A. Lind. Blogs, twitter, and breaking news: The produsage of citizen journalism. *Produsing theory in a digital world: The intersection of audiences and production in contemporary theory*, 80(2012):15–32, 2012.

[5] G. Fischer and C. Stevens. Information access in complex, poorly structured information spaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '91, pages 63–70, New York, NY, USA, 1991. ACM. ISBN 0-89791-383-3. doi: 10.1145/108844.108854. URL `http://doi.acm.org/10.1145/108844.108854`.

[6] N. Gershon and S. G. Eick. Visualization's new tack: Making sense of information. *IEEE Spectr.*, 32(11):38–40, 42, 44–7, 55–6, Nov. 1995. ISSN 0018-9235. doi: 10.1109/6.469330. URL `http://dx.doi.org/10.1109/6.469330`.

[7] I. Hargreaves and J. Thomas. *New news, old news.* Broadcasting Standards Commission, 2002.

[8] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE transactions on visualization and computer graphics*, 8(1):9–20, 2002.

[9] J. Ho and R. Tang. Towards an optimal resolution to information overload: An infomediary approach. In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '01, pages 91–96, New York, NY, USA, 2001. ACM. ISBN 1-58113-294-8. doi: 10.1145/500286.500302. URL `http://doi.acm.org/10.1145/500286.500302`.

[10] A. E. Holton and H. I. Chyi. News and the overloaded consumer: Factors influencing information overload among news consumers. *Cyberpsychology, Behavior, and Social Networking*, 15(11):619–624, 2012.

[11] H. S. Husin, J. A. Thom, and X. Zhang. Analysing user access to an online newspaper. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, pages 77:77–77:80, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3000-8. doi: 10.1145/2682862.2682875. URL `http://doi.acm.org/10.1145/2682862.2682875`.

[12] C.-Y. Lin and E. Hovy. Identifying topics by position. In *Proceedings of the fifth conference on Applied natural language processing*, pages 283–290. Association for Computational Linguistics, 1997.

[13] B. Nordenson. Overload! Journalism's battle for relevance in an age of too much information. *Columbia Journalism Review*, 47(4):30, 2008.

[14] I. Pentina and M. Tarafdar. From "information" to "knowing": Exploring the role of social media in contemporary news consumption. *Computers in Human Behavior*, 35:211–223, 2014.

[15] M. S. Pera and Y.-K. Ng. Utilizing phrase-similarity measures for detecting and clustering informative rss news articles. *Integrated Computer-Aided Engineering*, 15 (4):331–350, 2008.

[16] S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 120–123. IEEE, 2010.

[17] K. Purcell, L. Rainie, A. Mitchell, T. Rosenstiel, and K. Olmstead. Understanding the participatory news consumer. *Pew Internet and American Life Project*, 1:19–21, 2010.

[18] J. Rosen. National explainer: A job for journalists on the demand side of news. `http://archive.pressthink.org/2008/08/13/national_explain.html`, 2008. Accessed: 2/11/2016.

[19] D. M. Russell, M. Slaney, Y. Qu, and M. Houston. Being literate with large document collections: Observational studies and cost structure tradeoffs. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 3, pages 55–55. IEEE, 2006.

[20] A. G. Schick, L. A. Gordon, and S. Haka. Information overload: A temporal approach. *Accounting, Organizations and Society*, 15(3):199–220, 1990.

[21] H. Schütze. Introduction to information retrieval. In *Proceedings of the international communication of association for computing machinery conference*, 2008.

[22] D. M. Strong, Y. W. Lee, and R. Y. Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[23] Twitter Account: @GoogleTrends. Tweet: *+250% spike in "what happens if we leave the EU" in the past hour*. `http://twitter.com/GoogleTrends/status/746137920940056578`. Accessed: 27/10/2016.

[24] K. E. Weick, K. M. Sutcliffe, and D. Obstfeld. Organizing and the process of sensemaking. *Organization science*, 16(4):409–421, 2005.

[25] J. S. Yi, Y.-a. Kang, J. T. Stasko, and J. A. Jacko. Understanding and characterizing insights: How do people gain insights using information visualization? In *Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel evaLuation Methods for Information Visualization*, BELIV '08, pages 4:1–4:6, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-016-6. doi: 10.1145/1377966.1377971. URL `http://doi.acm.org/10.1145/1377966.1377971`.