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 @GoogleTrends Twitter account reported a 250% increase in searches for "What happens if we leave the EU?". 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.

Some months after Leonhardt's article was published, Journalism Professor Jay Rosen voiced his agreement with its premise in a blog post on the failure of journalism during the financial crisis; "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."[Rosen, 2008]

It has become apparent that prolific coverage alone is not enough to engage and support the public in understanding the complexities of 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 [2002] 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." [Nordenson, 2008, p.2]

In 2007, The Associated Press conducted an extensive field study 2008 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 [Holton and Chyi, 2012, Purcell et al., 2010, Fischer and Stevens, 1991] 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, 2008, p.37) from particular headlines, often found themselves reading the same summary-level content from different news sources. It was recommended that news providers support this by "designing innovative formats and creating easier pathways to deep content" [Associated Press and Context-Based Research Group, 2008, p.49]

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" [Schick et al., 1990, p.206]. Information overload is a multifaceted problem which can be modelled as a combination of three contributing factors (Figure 1.1).

These factors correspond directly to the three points previously identified from the Associated Press and Context-Based Research Group study. High information quantity leads to 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. [2010] further decompose 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 in quick succession over a short period of time.

Intuitively, the terms spatial and temporal quality seem illogically named, as a set of articles with high spatial quantity would cover a smaller area of information space and vice versa. High spatial quantity will therefore be refered to as *redundancy*, and high temporal quantity as *fragmentation*, since a sudden burst of articles published on the same topic suggests an unfolding story being told in parts.

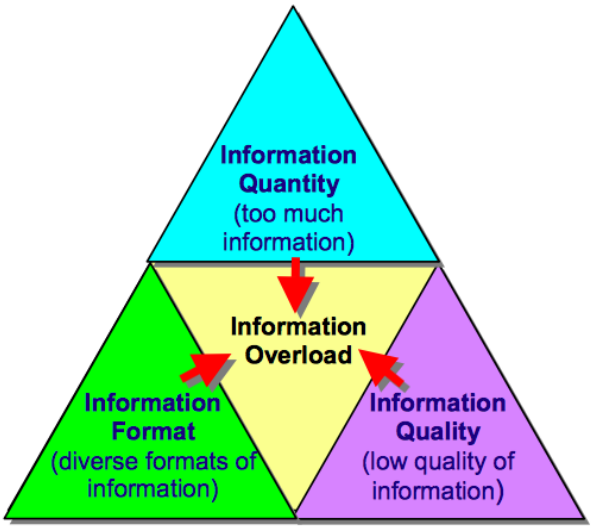


Figure 1.1: Dimensions of information overload, as defined by Ho and Tang [2001].

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

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.

Information Quality

In the context of factual data rather than news specifically, Strong et al. [1997] 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 [Strong et al., 1997] 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.

Fragmentation (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 content for the consumer to piece together in order 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 [Bruns et al., 2012].

Phuvipadawat and Murata [2010] 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 [Salton and Buckley, 1988] of the two messages and the number of named entities they have in common.

Redundancy (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 [Barzilay et al., 1999] 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 [Pera and Ng, 2008], 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. 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 four 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 [Pentina and Tarafdar, 2014]. 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, 'What's the story here?'" [Weick et al., 2005, p.85]

This relates directly to the news overload problem because sensemaking describes the contextual story resolution the Associated Press and Context-Based Research Group [2008] study identified news consumers are craving. 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 [Husin et al., 2014]. How then, do readers make sense of a collection of articles surrounding a particular topic?

When presented with a large document collection, [Russell et al., 2006] found all subjects began by clustering the contents into groups which formed a heuristic representation or mental model, used to provide an overview. However, current information infrastructure has been criticised for not supporting the cross-correlation between connected news articles [Rennison, 1994].

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."[Nordenson, 2008, p.10]

Similarly, Pentina and Tarafdar [2014] make 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 comprension of the data is information visualisation, or InfoVis. [Yi et al., 2008, Havre et al., 2002].

Yi et al. [2008] 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."[Gershon and Eick, 1995, p.39]

In addition to the insight that visualisation may be able to provide, there is evidence that visual representation metaphors also focus users' attention, support the learning process, and better utilise memory [Burkhard, 2004].

1.2 Towards Visualisation

The fact that news articles form a fairly narrow class of document is an advantage from a visualisation design perspective, due to the common elements they share. Articles published by commercial news producers typically contain:

- A headline
- A description, or *subhead*
- A publish date
- A category to which the article belongs

These attributes are useful for visualisation, since creating a spatial representation from text requires documents to be represented as vectors in high-dimensional feature space [Wise et al., 1995], and the presence of existing attributes makes articles more inherently comparable.

1.2.1 Mining Articles

The de-facto web format for news feed publishing is RSS (Rich Site Summary, or Really Simple Syndication.) The rise of the internet as a news platform has lead to many readers finding the most efficient method of reading news articles is to subscribe to various topic-specific news feeds and read what is automatically collated by their computers [Wang et al., 2006].

Although RSS–which is a subset of XML–is standardised[1], the practice of feed categorisation is not, meaning the granularity of topics which can be subscribed to is dependent on the publisher. This issue was addressed by Liu et al. [2009a], with the design of a system which could essentially split or join existing RSS feeds to synthesise new ones based on user-specified keywords and queries.

Despite its shortcomings, RSS remains the most universal option for accessing news feed content from a huge variety of news producers [O'Shea and Levene, 2011].

1.2.2 Chronology

Chronology is a vital aspect of news, and should clearly be preserved in any visualisation of news data as it provides a natural ordering to articles [Binh Tran, 2013].

[1] http://cyber.harvard.edu/rss/rss.html

Singh et al. [2015] designed a prototype for generating annotated timelines based on the Wikipedia entries long-running news stories. Use of Wikipedia rather than news feeds meant their document retrieval model was heavily dependent on Wikipedia's structure, but it also afforded a huge wealth of contextual information that made such detailed annotations possible.

Both ESTHETE [Goyal et al., 2013] and nReader [Wang et al., 2006] present timeline-centric views for collections of news articles based on underlying graphs of relationships between the articles. However, in both cases, the graph structure was not part of the final visualisation, so connections between entities were displayed in purely textual forms.

In contrast, ThemeRiver [Havre et al., 2002] introduces a novel view on topic frequency along the time axis to show thematic change over time within a collection of documents, similar to a smoothed histogram.

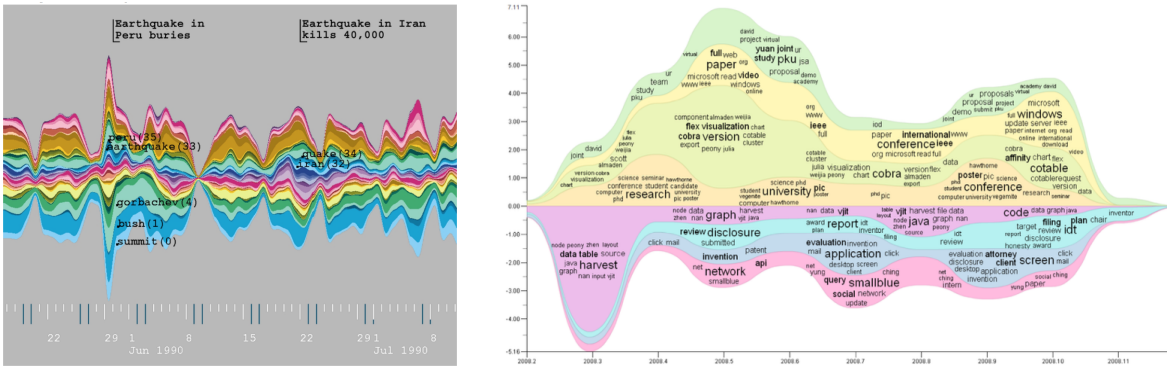


Figure 1.2: Similar visualisations from ThemeRiver [Havre et al., 2002] (left) and TIARA [Liu et al., 2009b] (right).

TIARA [Liu et al., 2009b] which cites ThemeRiver as an influencing design (see figure 1.2), displays a similar shaped graphical output but performs more detailed textual analysis, and displays related keywords in the output. Both visualisations support simple zooming and panning, but the interpolation required to soothe the histograms leads to inaccuracies on the frequency axis when users zoom in beyond a certain point.

1.2.3 Keyword Extraction

Extracting keywords from documents is not a new area of research. Various methods have been presented, the most well-known being TF-IDF (term frequency, inverse document frequency) [Salton and Buckley, 1988] which ranks the significance of a term t in a document d which belongs to a corpus C as follows:

$$\mathrm{TF}(t,d) = Occurrences(t,d) \div WordCount(d) \tag{1.1}$$

$$\mathrm{IDF}(t,C) = log_e\left(\frac{|C|}{|\{c \in C \mid t \in c\}|}\right) \tag{1.2}$$

$$\text{TF-IDF}(t,d,C) = TF(t,d) \times IDF(t,C) \tag{1.3}$$

The reliance on TF-IDF on a fixed background corpus results in a need to recompute the function for every document if any are added to or removed from the collection. This is impractical for large collections, and even in the case of large fixed collections it does not scale well, which has resulted in the development of other methods.

An approach derived from energy levels in quantum systems was proposed in [Carpena et al., 2009], where keywords were extracted based on their spatial distributions within a single text. The theory behind the approach is that typically, keywords occurrences are distributed in significant frequency clusters throughout a document, whereas non-relevant words are distributed with uniform frequency (see Figure 1.3).

This technique allows relevant keywords to be distinguished from non-relevant common words with similar total frequencies, without the use of a background corpus for comparison.

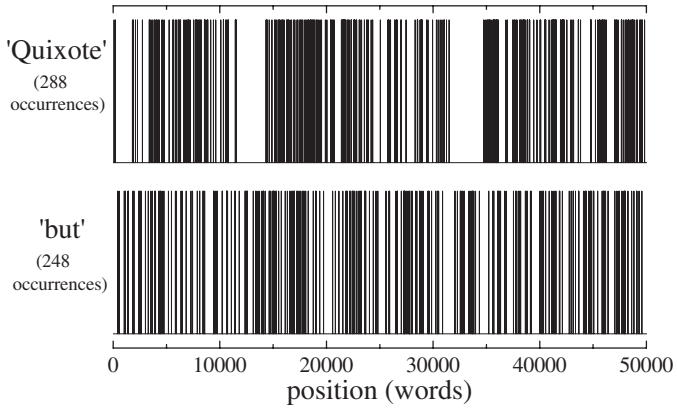


Figure 1.3: Frequency spectra of 'Quixote' (keyword, clustered distribution) and 'but' (non-relevant, uniform distribution) in the first 50,000 words of *Don Quixote*.

Several important observations have been made regarding keyword extraction for news articles specifically. Firstly, that important phrases in text are likely to be references to people, places and other named entities [Teitler et al., 2008]. Libraries such as the Stanford Named Entity Recognizer (NER) [Finkel and Manning, 2009] exist to extract these from text. Secondly, while 30% of an article's keywords are inferred and cannot be found within the text without intelligent input, 60% are present in the article's title and first few sentences [Lin and Hovy, 1997], since important facts are generally stated early.

1.3 The Metro Maps Metaphor

Significant work in the area of information visualisation–and in particular, information *cartography*–has been undertaken by Shahaf et al. [Shahaf and Guestrin, 2010, Shahaf et al., 2012a,b, 2013], in the domains of both news and science through the visualisation of article and journal data on metro maps.

The the metro map was introduced in [Shahaf et al., 2012a] to address the fact that previous timeline-based news summarisation systems could only represent simple linear stories; "In contrast, complex stories display a very non-linear structure: stories split into branches, side stories, dead ends, and intertwining narratives." [Shahaf et al., 2013, p.1]

Using the preexisting visual metaphor of a transit map–which a large number of people will already be familiar with–supports readers' comprehension, as it requires both time and effort for a reader to interpret visual metaphors which are new to them [Ziemkiewicz and Kosara, 2009].

Definition 1. *Metro Map [Shahaf et al., 2012a]: A metro map $\mathcal{M}$ is a pair (G, Π), where $G = (V, E)$ is a directed graph and Π is a set of paths, or metro lines in G. Each $e \in E$ must belong to at least one metro line.*

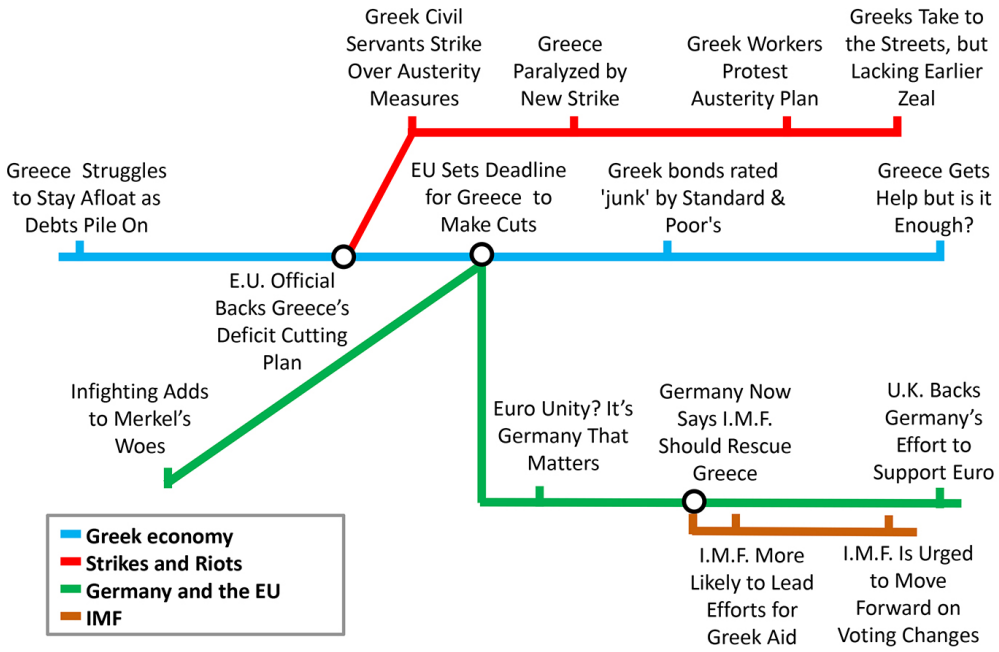


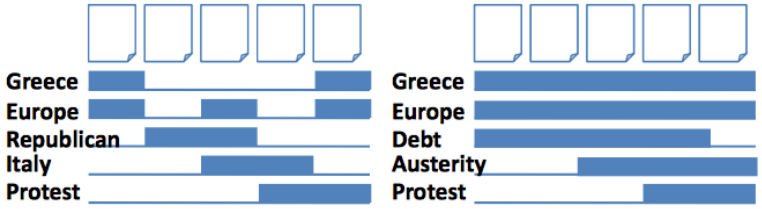
Figure 1.4: A metro map [Shahaf et al., 2012a] covering the Greek Debt Crisis.

A previously published method [Shahaf and Guestrin, 2010] for linking together chains of articles was discussed, and an objective function was created to formalise the characteristics of a 'good' metro map. The function defined was a composite based on three important characteristics, all of which are broadly applicable to the visualisation of any similar corpora; coherence, coverage, and connectivity.

1.3.1 Coherence

Let $\mathcal{D}$ be a set of articles, and $\mathcal{W}$ be a set of words or phrases, such that each article is a subset of $\mathcal{W}$. A *coherent* chain of articles through $\mathcal{D}$ is one where transitions between documents are smoothed by common overlapping keywords from $\mathcal{W}$, creating a better narrative flow [Shahaf and Guestrin, 2010] as depicted in Figure 1.5.

Coherence, intuitively, seems to be closely linked to idea of story resolution detailed in Section 1.1.1. This presents a question which could later be explored further; does forming coherent chains of articles provide the story resolution that participants in the Associated Press study were so desperately seeking from current events journalism?



The titles of the articles which made up the two chains were as follows:

Chain A (left)	Chain B (right)
Europe weighs possibility of debt default in Greece	Europe weighs possibility of debt default in Greece
Why Republicans don't fear a debt default	Europe commits to action on Greek debt
Italy; The Pope's leaning toward Republican ideas	Europe union moves towards a bailout of Greece
Italian-American groups protest 'Sopranos'	Greece set to release austerity plan
Greek workers protest austerity plan	Greek workers protest austerity plan

Figure 1.5: An incoherent chain with jittery transitions between topics (Chain A, left) alongside a more coherent chain of articles (Chain B, right). A bar corresponds to the presence of a word in the article above it. [Shahaf et al., 2012a]

1.3.2 Coverage

As in the previous section, let $\mathcal{D}$ be a set of articles, and $\mathcal{W}$ be a set of words or phrases of which the articles are composed. The coverage function for a word in a given document $d_i \in \mathcal{D}$ specified in Equation 1.4 can be quantified using any measure of how well d_i covers w, for example TF-IDF$(w, d_i, \mathcal{D})$ (See Equation 1.3) [Shahaf et al., 2012a].

$$cover_{d_i}(w) : \mathcal{W} \rightarrow [0, 1] \tag{1.4}$$

Extending the notion of coverage to maps–which can be abstracted to sets of documents–introduces the idea of *diversity*. If a map already contains documents which for a sufficient coverage for some word w, then there is nothing to be gained by adding another document to $\mathcal{D}$ which has high coverage of w alone. This relates back to the principles of spacial and information quality discussed in Section 1.1.1, especially the importance of value-added by every individual document in a collection. In this case, maps which cover a maximal number of $w \in \mathcal{W}$ should be preferential. A simple additive definition for map coverage such as Equation 1.5 [Shahaf et al., 2012a] would not reward this kind of diversity.

$$cover_{\mathcal{M}}(w) = \sum_{d_i \in docs(\mathcal{M})} cover_{d_i}(w) \tag{1.5}$$

Therefore, an alternative definition for map coverage was chosen, which will not increase significantly if another document which covers an already covered feature is added to $\mathcal{D}$ (Equation 1.6 [Shahaf et al., 2012a]).

$$cover_{\mathcal{M}}(w) = 1 - \prod_{d_i \in docs(\mathcal{M})} (1 - cover_{d_i}(w)) \tag{1.6}$$

Finally, the definition of map coverage is extended to the coverage of the corpus $\mathcal{D}$, rather than just single features. If each feature is weighted, according to frequency, then for each $w \in \mathcal{W}$ we have some λ_w. The coverage of a corpus $\mathcal{D}$ by a metro map $\mathcal{M}$ can then be defined as in Equation 1.7 [Shahaf et al., 2012a].

$$Cover(\mathcal{M}, \mathcal{D}) = \sum_{w \in \mathcal{W}} \lambda_w cover_{\mathcal{M}}(w) \tag{1.7}$$

1.3.3 Connectivity

The final property is the most simply defined; the connectivity of a metro map is the number of paths in Π which intersect [Shahaf et al., 2012a].

$$Connectivity(\mathcal{M}) = \sum_{i<j} \mathbb{1}(p_i \cap p_j \neq \emptyset) \tag{1.8}$$

1.3.4 Limitations

Perhaps the biggest limitation of the system developed in [Shahaf et al., 2012a, 2013] is the nature of the corpus $\mathcal{D}$; it is a fixed dataset, meaning all a user can do is query it for certain non-recent events with no way of specifying a different corpus themselves. From a historical reference perspective the output provided for certain queries is interesting, but it is not possible to use the system as a replacement to a news feed aggregator or similar tool, which seems as though it would be the next logical step.

1.4 Evaluation Methods

Shahaf et al. [2012a] evaluated their system both for accuracy and with a user study. The accuracy evaluation tested whether the system included the most 'important' (as decided by experts) documents in the map. The user study focused on the strength of the results returned by specific queries, where output was transformed into a structureless list in order for the study to be double-blind against the other methods. The evaluation was performed between-subjects, so background knowledge had to be controlled for. Output was compared with that from Google News and a TDT (topic detection and tracking) method presented in [Nallapati, 2003]. This approach to evaluation is less relevant to my proposed system, since it was actually evaluating the performance of the system in selecting documents based on a query, rather than visualising the documents on a map. In contrast, a visualisation and its usability is precisely the aspect of my process which I would ultimately need to evaluate.

Andrews [2006] found that the evaluation of measures of usability such as task completion time and effectiveness can only be accurately conducted as part of a summative formal experiment. This is because formative tests such as think-aloud experiments require users to alter their behaviour and leads to slower actions [Ericsson and Simon, 1980].

The evaluation of TIARA [Liu et al., 2009b] was a more relevant method than the Metro Map evaluation as it was conducted against a baseline system which did not share any of its advanced features, although it was tailored for the same task; email analysis. A series of questions were asked of participants, who used either TIARA or the baseline system to answer. The response time and accuracy of the participants was recorded, as well as their levels of satisfaction after completing the task. This evaluation used between-subjects designs and therefore required the use of a different dataset for each task, as the nature of the sensemaking means any repetition of the evaluation task on the same data would see participants' performance improve significantly due to recall alone.

1.5 Summary

In summary, this review began with an exploration of the issue of news overload using the findings of [Associated Press and Context-Based Research Group, 2008], a field study conducted by the Associated Press into the news consumption habits of young people. The findings of the study were then discussed, firstly in the context of four dimensions of information overload [Ho and Tang, 2001, Bergamaschi et al., 2010] and secondly in terms of how they relate to sensemaking, the cognitive process this project aims to support.

Information visualisation was identified as one common approach to both support sensemaking and reduce information overload, so various methods for visualising text-based documents and news articles specifically were presented and compared [Goyal et al., 2013, Havre et al., 2002, Liu et al., 2009b, Singh et al., 2015, Wang et al., 2006] as well as methods for transforming news articles into feature vectors for visualisation, such as mining, entity recognition, and keyword extraction.

Lastly, the work of Shahaf et al. [Shahaf and Guestrin, 2010, Shahaf et al., 2012a,b, 2013] which is particularly relevant to the aims and objectives of this project was discussed in detail, with an explanation of key concepts (*coherence*, *coverage* and *complexity*) defined in [Shahaf et al., 2012a] which are applicable to all graph-based representations of document collections.

Conducting this research has resulted in a clear direction of focus and scope for my project, and an understanding of what I could contribute to the domain of visualisation-based solutions to information overload. My aim is to integrate user-specified news feeds into the approach outlined in [Shahaf et al., 2012a] to generate custom maps based on current events, with a richer interactive display format and the addition of other techniques for reducing information load such as removing subsumed articles from the model [Pera and Ng, 2008] and attempting to provide contextual background in the visualisation itself [Singh et al., 2015].

The resultant system would be a news feed aggregator with visually structured output, and to the best of my knowledge would be the first of its kind.

Bibliography

K. Andrews. Evaluating information visualisations. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, pages 1–5, New York, NY, USA, 2006. ACM. ISBN 1-59593-562-2. doi: 10.1145/1168149.1168151. URL `http://doi.acm.org/10.1145/1168149.1168151`.

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.

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.

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

G. Binh Tran. Structured summarization for news events. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 343–348, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2038-2. doi: 10.1145/2487788.2487940. URL `http://doi.acm.org/10.1145/2487788.2487940`.

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.

R. A. Burkhard. Learning from architects: the difference between knowledge visualization and information visualization. In *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*, pages 519–524. IEEE, 2004.

P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. Coronado, and J. Oliver. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E*, 79(3):035102, 2009.

K. A. Ericsson and H. A. Simon. Verbal reports as data. *Psychological review*, 87(3):215, 1980.

J. R. Finkel and C. D. Manning. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 141–150. Association for Computational Linguistics, 2009.

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.

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`.

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

R. Goyal, R. Malla, A. Bagchi, S. Mehta, and M. Ramanath. Esthete: A news browsing system to visualize the context and evolution of news stories. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 2529–2532, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. doi: 10.1145/2505515.2508208. URL `http://doi.acm.org/10.1145/2505515.2508208`.

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

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.

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`.

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.

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.

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.

B. Liu, H. Han, T. Noro, and T. Tokuda. Personal news rss feeds generation using existing news feeds. In *International Conference on Web Engineering*, pages 419–433. Springer, 2009a.

S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. Interactive, topic-based visual text summarization and analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 543–552, New York, NY, USA, 2009b. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646023.

R. Nallapati. Semantic language models for topic detection and tracking. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL*

2003 student research workshop-Volume 3, pages 1–6. Association for Computational Linguistics, 2003.

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

M. O'Shea and M. Levene. Mining and visualising information from rss feeds: a case study. *International Journal of Web Information Systems*, 7(2):105–129, 2011.

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.

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.

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.

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.

E. Rennison. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, UIST '94, pages 3–12, New York, NY, USA, 1994. ACM. ISBN 0-89791-657-3. doi: 10.1145/192426.192429. URL `http://doi.acm.org/10.1145/192426.192429`.

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.

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.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

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

D. Shahaf and C. Guestrin. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–632. ACM, 2010.

D. Shahaf, C. Guestrin, and E. Horvitz. Trains of thought: Generating information maps. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 899–908, New York, NY, USA, 2012a. ACM. ISBN 978-1-4503-1229-5. doi: 10.1145/2187836.2187957. URL `http://doi.acm.org/10.1145/2187836.2187957`.

D. Shahaf, C. Guestrin, and E. Horvitz. Metro maps of science. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1122–1130, New York, NY, USA, 2012b. ACM. ISBN 978-1-4503-1462-6. doi: 10.1145/2339530.2339706. URL `http://doi.acm.org/10.1145/2339530.2339706`.

D. Shahaf, J. Yang, C. Suen, J. Jacobs, H. Wang, and J. Leskovec. Information cartography: creating zoomable, large-scale maps of information. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105. ACM, 2013.

J. Singh, A. Anand, V. Setty, and A. Anand. Exploring long running news stories using wikipedia. In *Proceedings of the ACM Web Science Conference*, WebSci '15, pages 57:1–57:2, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3672-7. doi: 10.1145/2786451.2786489. URL `http://doi.acm.org/10.1145/2786451.2786489`.

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

B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: a new view on news. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 18. ACM, 2008.

T. Wang, N. Yu, Z. Li, and M. Li. nReader: Reading News Quickly, Deeply and Vividly. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 1385–1390, New York, NY, USA, 2006. ACM. ISBN 1-59593-298-4. doi: 10.1145/1125451.1125707.

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

J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the 1995 IEEE Symposium on Information Visualization*, INFOVIS '95, pages 51–, Washington, DC, USA, 1995. IEEE Computer Society. ISBN 0-8186-7201-3.

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`.

C. Ziemkiewicz and R. Kosara. Preconceptions and individual differences in understanding visual metaphors. In *Computer Graphics Forum*, volume 28, pages 911–918. Wiley Online Library, 2009.