Identifying Archetypal Perspectives in News Articles

Chris P. Bowers
The School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK
cpb@cs.bham.ac.uk

Russell Beale
The School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK
rxb@cs.bham.ac.uk

Robert J. Hendley
The School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK
rjh@cs.bham.ac.uk

A novel approach to news aggregation is proposed. Rather than ranking or summarisation of cluster topics, we propose that articles are grouped by topic similarity and then clustered within topic groups in order to identify archetypal articles that represent the various perspectives upon a topic. An example application is examined and a preliminary user study is discussed. Future applications and evaluation of validity are outlined.

News aggregation, perspective archetypes, keyphrase extraction, information retrieval

1. INTRODUCTION

The growth in communication technologies has resulted in an explosion in the number of sources providing news, blogs, articles, reviews and opinions on a wide range of topics. This increase in digital content has created a need for effective indexing, categorisation and retrieval of news articles. The vast majority of news organisations provide content in a digital format and announce new articles in the form of a feed. Feeds are a standardised format, most commonly Really Simple Syndication (RSS).

A popular way to access articles on the web is through aggregators. These services gather feeds together and apply filtering. Which news articles are chosen and how the feeds are filtered varies from one service to another. Early aggregation services operated manually, requiring a reader, editor or author to make a conscious decision about whether to include some news article or not. Most are now almost fully automated, gathering, filtering and indexing news articles from thousands of sources e.g. Google News[1]. Some use crowd-sourcing to gather and rank articles from a large user base such as digg[2] or reddit[3] .

The main objective of a news aggregation service is to provide the user with a single entry point for news articles that are likely to be of interest in an easily accessible form. To achieve this, a news aggregator, whether automatic, crowd sourced or manual, must truncate the number of articles presented to the user. In most cases this is done with some form of ranking. However, with each new article comes a potential new viewpoint on a topic. With news aggregators that return ranked results it is difficult to get an overview of how current opinion varies over a range of topics.

An alternative approach is to group articles that contain similar or related news. The challenge with grouping articles in this way is how to reduce the clusters into an overview that is easily accessible. One might try to find an archetype article from each topic cluster (Leuski 2001) or form some human readable summarisation of each cluster (McKeown et al 2002). Previous work has attempted to automatically generate a multi-document summary of news articles using natural language processing or sentiment analysis techniques (Godbole et al 2007; McKeown et al 2002). In practice, although potentially very powerful, these systems are complex, difficult to implement in a robust way and tend to abstract and generalise the various perspectives on a topic (Liu 2010).

In this work we discuss an alternative approach to presenting topic cluster to users. Rather than attempt to generate a summarisation of a set of related news articles, we propose that related articles be filtered using a statistical process in an attempt to identify the varying perspectives on a topic using a vector space document model. The novelty in the proposed approach comes not from the grouping

[1] http://news.google.com/
[2] http://digg.com/
[3] http://www.reddit.com/

of news articles by topic similarity but from the attempt to cluster within these groupings to identify the differences in perspective upon that topic. By clustering articles within a topic group it is possible to identify archetypical articles that epitomise the various viewpoints expressed within a related set of articles.

This approach has a number of advantages over summarisation of topic clusters. It provides an accessible overview of a balanced set of perspectives and opinions on a particular topic. This might allow users to explore the range and extremities of opinion on a particular topic and could be extended to support hierarchies allowing users to drill down to more specific perspectives whilst still retaining an overview of the current articles related to the overall topic. Similarly, single-linkage clustering could be used to visualise the relationship between perspectives amongst a set of articles in a topic cluster. This would be a powerful tool for researchers/journalists.

This approach could be applied to a variety of domains besides news aggregation such as presenting search results or identifying perspectives circulating in the blogosphere/twitterspace about an organisation, individual or product.

2. GENERATING ARCHETYPAL PERSPECTIVES

2.1. A vector space model of news articles

There are many approaches to automatically abstracting the content of a text document. Typically the most robust methods utilise a bag-of-words vector space model. For the purposes of this work we utilise the term frequency-inverse document frequency (TF-IDF) weighting to determine the relevance of a given phrase to a documents content (Salton et al 1988). The algorithm works as follows:

- A set of news articles are retrieved using a pre-determined list of news sources (in the form of RSS feeds).
- Each article is transformed into a bag-of-words and each word is stemmed to remove inflection and derivations and the remaining words are stored with their frequency of occurrence for each article. Stop words are filtered out to improve computational performance.
- The TF-IDF value for each word is then calculated (Salton et al 1988). Higher values occur when a word is used frequently within a document but not in other documents. In one sense, it measures how surprising it is to see that word featured as often as it does (Beale 2007).
- Phrases consisting of more than one word can also be considered. Groupings of words can be considered to be a phrase if they appear adjacent and in the same order with a high frequency within an article. This process can be generalised to find longer phrases but we restrict the extraction to single word or two word phrases. TF-IDF values can then be calculated for these multi-word phrases in the same way as for single-word phrases.

The content of an article can therefore be summarized using a vector of the TF-IDF weight for each phrase. Conversely, a phrase vector can be used to summarise how a phrase relates to each article using the TF-IDF values for that phrase for each article.

2.2. Grouping by topic

Given that phrase vectors represent the content of documents, some of these are more aligned to the various topics of the articles than others. A simple ranking of words and phrases over all articles is performed using the magnitude of the phrase vector. Those phrases with the greatest magnitude are considered to be the most descriptive of the content of the articles. If the user selects one, or a combination, of these phrases, a set of relevant articles can be retrieved.

2.3. Clustering by perspective

Given a set of articles described by vectors their similarity to each other can be determined by their relative direction in the vector space. Those documents that are similar should result in vectors that point in a similar direction. This similarity can be estimated using the Cosine similarity function.

Using this similarity measure, articles are then clustered using an evolutionary clustering approach (Hruschka et al 2009). An evolutionary algorithm is used to find optimal cluster membership by simultaneously maximising the similarity within clusters and minimising the similarity between clusters (Davies & Bouldin 1979).

In this work we consider perspective, in a very general sense, to describe diverging viewpoints on a topic. Most clustering applied to text analysis is in the form of topic grouping where the objective is to group articles that discuss a common topic. In other words emphasis is placed on identifying similarities between documents. However, relatively little work has investigated the ability to distinguish between perspective where the emphasis is on distinguishing between documents on the same topic.

Table 1: *Agencies used to source online news articles*

	Agency	Type
1	Google News	News aggregator
2	BBC News	Broadcast news
3	Fox News	Broadcast news
4	Daily Mail	Daily tabloid
5	Channel 4	Broadcast news
6	The Economist	Weekly intl. news
7	Birmingham Mail	Regional tabloid
8	The Metro	Daily tabloid
9	The Sun	Daily tabloid
10	Reuters	Intl. news agency
11	Sky News	Broadcast news
12	The Scotsman	Regional broadsheet
13	news.com.au	News aggregator
14	NDTV News	Broadcast news
15	Al Jazeera	Broadcast news
16	The Telegraph	National broadsheet
17	The Huffington	Post Intl broadsheet
18	The Guardian	National broadsheet

Lin and Hauptmann (2006) have shown that differences in ideological perspective can be successfully detected from a pair of documents using a simple bag-of-words model.

2.4. Identifying archetypes

Given a set of clusters, which represent the various perspectives on a topic, an archetype is used to characterise each cluster. In this case archetypes are chosen by finding the article, within each cluster, which is closest to the cluster centroid.

3. PRELIMINARY EVALUATION

Table 1 lists a sample of news agencies that offer freely available RSS feeds of their top stories updated on a regular basis.

Using these news sources 354 separate news articles were retrieved. From these articles 2919 distinct phrases were extracted from the titles and brief description provided by the RSS feeds. Each of these phrases was then ranked by TF-IDF value and table 2 defines the top 10 key phrases obtained.

For this example we choose two of the highest ranking terms (highlighted in Table 2). A total of 24 articles are associated with these phrases. Table 3 shows the source, title and resulting clustering of each of these articles. For this example, eight

Table 2: *Top 10 key phrases describing the content of retrieved new feeds*

Phrase	# of associated articles	TF-IDF vector magnitude
Police	41	1.40
Tax	15	1.37
Nick Clegg	15	1.26
Murder	21	1.17
Libya	16	1.07
Gaddafi	15	1.03
Ivory Coast	14	1.02
Coast	16	0.97
Government	29	0.95
Warning	14	0.93
...	...	...

clusters were used with the archetypal article for each of these clusters highlighted.

Some insight into the performance of the archetypal extraction can be gained from looking at the archetypes in this example from a human rather than an algorithmic perspective. For example cluster 2 consists of articles discussing issues with NATO air strikes and the alleged use of civilians as human shields. Cluster 4 primarily consists of articles discussing planned attacks on civilians. The remaining clusters are much smaller but each consists of one or more distinct viewpoints on the topic.

3.1. User Study

In order to gain a greater understanding of how people identify various perspectives on a topic a small participatory study was undertaken. Each participant was provided with the same articles, outline in table 3, printed onto cards and asked to do the following:

- Sort the articles into clusters corresponding to different perspectives on the topic. No restrictions were defined on the number of clusters or the number of articles that a cluster can contain.
- Provide a brief summary of each cluster.
- Choose a single article from each cluster which best exemplifies that cluster.

A total of 6 participants (2 females) undertook the study and were handed the cards at the beginning of the experiment in a randomly order pile. Two primary methods of clusters were observed: (1) The participant spread all the cards down on the table

Table 3: *Clustering and identified archetypes for a set of articles from a topic group*

Agency	Article Title	ID
Channel 4	Libya ready for political reform says Gaddafi spokesman	0
Channel 4	Gaddafi regime seeking an 'end to the conflict'	0
Fox News	Groups Rip U.N. Adviser Over Qaddafi Rights 'Prize'	1
Google News	Libya rebels accuse NATO of holding back in fight against Gadhafi...	2
Channel 4	Libyan rebels: NATO failing civilians	2
Reuters	Libya rebels retake land: NATO cites air strike woes	2
The Scotsman	Libya: Gaddafi uses human shields to stop air strikes	2
The Telegraph	Libyan rebel leader condemns NATO inaction	2
The Guardian	Gaddafi forces using civilians as human shields: says France	2
Sky News	Gaddafi Sends Message To Barack Obama	2
BBC News	RAF jets join Libya ground attack	2
Daily Mail	Libya: Do we have the stomach to keep bombing Gaddafi for 6 mon...	3
The Guardian	Blessed are the wagers of just war? — Mischa Geracoulis	3
NDTV News	Libya photos show abuses under Gaddafi	4
BBC News	Libya 'planned to kill civilians'	4
The Telegraph	Footage of 'government attack on Brega'	4
Channel 4	Libya: coalition air strike kills rebels near Brega	4
The Guardian	Brega hit by Gaddafi forces' shells - dramatic amateur video footage	4
The Guardian	Gaddafi backers demolish rebel mosque	4
The Telegraph	Q&A: What's at stake in Ivory Coast	5
The Guardian	Ouattara forces 'storming Gbagbo's bunker'	5
Channel 4	Libya: Gaddafi forces loosen rebel grip on Misrata	6
The Metro	Col Gaddafi's son says Libya defector Musa Kusa should not be trust..	7
BBC News	'No deal' with Musa Kusa: Hague	7

so that each could be seen allowing them to cluster whilst maintaining an overview of all articles. (2) Alternatively, the participant went through the pile of articles one at a time placing them into clusters.

In the latter case participants tended to reconsider their clustering, often breaking clusters up into sub clusters or combining clusters. This is analgous to hierarchical clustering. Therefore, hierarchical clusters of perspective may prove to be an intuitive structure for users to manipulate.

Interestingly, there was little consistency in the assignment of clusters and summations. Participants tended to have very different opinions on what constituted a differing perspective. Some chose political, social, or emotive themes, whilst others chose to cluster on more factual points. Its not surprising that perspective is very subjective and this highlights the difficulty in evaluating automated perspective gathering with respect to human performance. However, when shown the clusters identified by the automated approach, participant feedback was generally positive.

4. DISCUSSION

As a work in progress we have demonstrated that extracting archetypes of perspective on a topic automatically from news archives and feeds is plausible. However, perspective is cleary subjective and this may impact on the usefulness of an automated approach to indentifying perspectives.

This prelimanary study is limited and there are a number of ways in which this work could be significantly improved.

The phrase extraction mechanism could be extended to use more powerful approaches which account for synonyms etc. More sophisticated representations used for sentiment analysis may account for the syntactic and semantic structure within the text. This might result in more accurate identification of perspectives but is likely to increase both computational costs and data requirements (Renz et al 2003). Similarly, explicit analysis of affect (Crawford & Henry 2004; Smith et al 2007) could

be used to further distinguish articles based upon the authors perspective on the topic.

The vector based evolutionary clustering approach used in this work was chosen because it is relatively robust (Hruschka et al 2009). However, there is no guarentee that it is optimal so when larger and more complex scenarios are taken into account then the clustering should be more considered. In this example the optimal number of centroids was discovered by user trial and error and this would need to be addressed in future work.

4.1. Future Evaluation

In order to assess the validity of this approach a more rigorous evaluation, in the context of existing news aggregation systems, is required. Surprisingly little work has been undertaken to understand the role of automated information retrieval processes for new aggregation in terms of usability and impact on user performance. Those existing studies are based on document summarisation approaches (Mckeown et al 2005). Typically these are assessed using a quantitative method which compares generated summaries to some hand crafted gold standard summary or other user-independent approaches (Bogers & van den Bosch 2007)]. However, this form of assessment often struggles to examine the generalisability of an approach. A preferred alternative is to use an information retrieval task, completed by human participants, where performance is measured by how quickly and accurately users retrieve some required information (Jing et al 1998).

The next step for this work is to undertake a comparative user study against state-of-the-art news article summarisation and ranking approaches. The study will be task-based requiring users to answer questions related to facts presented within a set of news articles. We will use a post task questionnaire to produce satisfaction ratings and assess the perceived success in providing an overview of perspectives on a topic.

From our preliminary study it is clear that there are associated issues with generating perspectives related to the subjective nature of perspective. We should evaluate the ability of our proposed system to truly represent the various viewpoints expressed in a set of articles and how users perceive bias in the presented archetypes. It is difficult to define an objective measure which can be calculated automatically. Therefore any evaluation is likely to be subjective. For example, one approach might be to compare archetypes produced by the system against those selected from the same set of articles by users.

The robustness of the approach should also be tested to ensure that the system works well in a wide variety of cases.

REFERENCES

Beale, R. (2007) *Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing*. Intl. J. Human Computer Studies, 65 (5):421-433.

Bogers, T. & van den Bosch, A. (2007) *Comparing and Evaluating Information Retrieval. Algorithms for News Recommendation*. In Proceedings of the 2007 ACM conference on Recommender systems (RecSys '07). ACM, New York, NY, USA, 141-144.

Crawford, J.R. & Henry, J. D. (2004) *The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normal data in a large non-clinical sample*. J. Clinical Psychology, 43:245-265.

Davies, D.L. & Bouldin, D.L. (1979) *A cluster separation measure*. IEEE Trans. Pattern Anal. Machine Intell. 1 (4). 224-227

Godbole, N., Srinivasaiah, M. & Skiena, S. (2007) *Large-scale sentiment analysis for news and blogs*. In proceedings of ICWSM.

Hruschka, E.R., Campello, R.J.G.B., Freitas, A.A. & Carvalho, A.C.P.L.F. (2009) *A survey of evolutionary algorithms for clustering*. IEEE Trans. on Systems, Man and Cybernetics, Part C: Applications and Reviews 39(2):133-155

Jing, H., Barzilay, R., Mckeown, K. & Elhadad, M. (1998) *Summarization Evaluation Methods: Experiments and Analysis*. In proc. AAAI Symposium on Intelligent Summarization, 60-68

Leuski, A. (2001) *Evaluating document clustering for interactive information retrieval*. In Proc. 10th international conference on Information and knowledge management (CIKM '01), Henrique Paques, Ling Liu, and David Grossman (Eds.). ACM, New York, NY, USA, 33-40.

Lin, W.H., & Hauptmann, A. (2006) *Are these documents written from different perspectives? A test of different perspectives based on statistical distribution divergence.* In Proc. 21st Intl. Conf. on Computational Linguistics (ACL-44). Association for Computational Linguistics, pp.1057-1064

Liu, B. (2010) *Sentiment analysis: a multifaceted problem*. IEEE Intelligent Systems, 25, pp. 7680

McKeown, K. R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L. , Nenkova, A., Sable, C., Schiffman, B. & Sigelman, S. (2002) *Tracking and summarizing news on a daily basis with columbias newsblaster*. In Proc. 2002 Human Language Technology Conference (HLT'02). Morgan Kaufmann Publishers Inc., pp.280-285.

Mckeown, K., Passonneau, R.J., Elson, D.K., Nenkova, A. & Hirschberg, J. (2005) *Do Summaries Help? A Task-Based Evaluation of Multi-Document Summarization*. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05). ACM, New York, NY, USA, 210-217.

Renz, I., Ficzay, A. & Hitzler, H. (2003) *Keyword Extraction for Text Characterization*. In Proc. 8th Intl Conf. on Application of Natural Language to Information Systems, GI (2003), 228-234.

Salton, G. & Buckley, C. (1988) *Term-Weighting approaches in Automatic Text Retrieval*. Information Processing & Management, 24(5):513-523.

Smith, C.J., Rumbell, T., Barnden, J.A., Hendley, R.J., Lee, M.G., Wallington, A.M. & Zhang, L. (2007) Don't worry about metaphor: affect detection for conversational agents. Proc. 45th meeting of the Association for Computational Linguistics, ACL, pp. 37-40.