Analysis of Open Answers to Survey Questions through Interactive Clustering and Theme Extraction

Fredrik Espinoza, Ola Hamfors, Jussi Karlgren, Fredrik Olsson, Per Persson, Lars Hamberg,
Magnus Sahlgren
Gavagai
Stockholm, Sweden

ABSTRACT

This paper describes design principles for and the implementation of Gavagai Explorer—a new application which builds on interactive text clustering to extract themes from topically coherent text sets such as open text answers to surveys or questionnaires.

An automated system is quick, consistent, and has full coverage over the study material. A system allows an analyst to analyze more answers in a given time period; provides the same initial results regardless of who does the analysis, reducing the risks of inter-rater discrepancy; and does not risk miss responses due to fatige or boredom. These factors reduce the cost and increase the reliability of the service. The most important feature, however, is relieving the human analyst from the frustrating aspects of the coding task, freeing the effort to the central challenge of understanding themes.

Gavagai Explorer is available on-line at http://explorer.gavagai.se

CCS CONCEPTS

•**Information systems** → **Clustering;** *Online analytical processing;* •**Human-centered computing** → User interface design;

ACM Reference format:
Fredrik Espinoza, Ola Hamfors, Jussi Karlgren, Fredrik Olsson, Per Persson, Lars Hamberg, Magnus Sahlgren. 2018. Analysis of Open Answers to Survey Questions through Interactive Clustering and Theme Extraction. In *Proceedings of Conference on Human Information Interaction & Retrieval, New Brunswick, NJ, USA, March 11–15, 2018 (CHIIR'18),* 4 pages.
DOI: http://dx.doi.org/10.1145/3176349.3176892

OPEN ANSWERS TO SURVEYS

Open answers in surveys and questionnaires are a challenge for analysts: how to report the collected responses together with more quantitative data elicited from respondents is not obvious. Typically a team of human analysts have been given the text responses together with a manually determined coding scheme, discussed and revised at intervals. The task of the analysts is to label the responses according to the coding scheme and to extract samples from the responses to anchor the labels in the data. In example (1) some extracts from reviews for a hotel are shown.

This coding procedure, converting the open responses into a structured form, requires time and expertise on the part of the analyst, both of which come at a cost. The effort involved in coding open answers is simultaneously intellectually non-trivial and demanding, but still monotonous: analyst fatigue and frustration risks leading to both between-analyst and within-analyst inconsistencies over time in reporting. This challenge is well-established both in the market research field and in scientific studies.[1] It takes about 1 minute for a human to categorise an abstract[2], when the categories are already given. If the task is to explore a set of responses and define and revise categories or labels as you go it will involve more effort and require more time per item.

(1) a. I would definately recommend this hotel, the location was great!
b. Had I known, I would NOT have chosen this hotel for my busy work visit in which I needed quiet time in hotel to do work.
c. Modern, stylish hotel with numerous, pretty decent restaurants in the area!

This paper describes a productivity tool for interactive coding, i.e. exploring and assigning thematic labels to open responses, based on a back-end technology which learns terminology and semantic relations from text.[3]

USE CASE

The purpose of including open questions in a survey is to explore the underlying motivations of the respondents with respect to some topic of interest. These motivations can be known in advance, they may be somewhat predictable, or they may be entirely unknown to the researcher. The resulting analysis, which is intended to give insights form the basis of e.g. strategic market decisions or other actions for the client, will be a set of such themes, with relevant quotes extracted from responses, reported together with their relative strengths and quantitative statistics on the numbers of respondents involved in discussing each theme.

The ambiguity, vagueness, and fluidity over time of human vocabulary is often described as a problem. This perspective does not do justice to the nature of human communication. The adaptability of human vocabulary and thus the entire human communication system is useful: it allows new terms to be coined, established terms

CHIIR'18, New Brunswick, NJ, USA

ISBN 978-1-4503-4925-3/18/03.
DOI: http://dx.doi.org/10.1145/3176349.3176892

[1] E.g. O'Cathain and Thomas [9] and many others.

[2] As shown by e.g. Macskassy et al. [7], McCallum et al. [8], Schohn and Cohn [14].

[3] This approach builds on a long-standing strand of research in information retrieval which builds on interplay between similarity based clustering and end-user assessment of clusters, such as Cutting et al. [1], Jardine and van Rijsbergen [6], Pirolli et al. [10], Sanderson and Croft [13] and many others.

to be recruited into service ad hoc to fit the needs of some discourse, and various discourses to be associated or contrasted through term choice. The challenge for the analyst of our specific use case is in fact exactly the reason why open answers are useful: if the choice of words were entirely predictable, the information captured through open answers would be so much less rich and valuable.

A question on trustworthiness of text gave the answers given in (2), referring to the various qualities the readers take into account. [4]

(2)
a. The appearance of the text, the quality of its design and polish.
b. How enjoyable and fun it is, how it addresses its readers, and who has written it.
c. Who wrote it and why.
d. Does it speak to me?

There are at least two themes in these four responses: the *source* of the text and the *audience design* of the text. The first theme was an expected theme, the second somewhat unexpected, and it would have been difficult for an editor to instruct a coding scheme to make note of terms such as *speak*, and *address* before the fact.

This sort of information is exactly what the study was designed to find. The intention underlying the design presented here is to empower the analyst to fold together Xs and Ys into a topically coherent theme, retaining the variation found in the material, not to normalise the behaviour of the respondents into a uniform vocabulary given before the responses.

INTERACTIVITY, NOT AUTOMATION

Our design principles are based on human language being useful as is, and on automating drudgery, not creation of insights.

Design principle 1: Empowering analysts, not replacing them. A repetitive and frustrating task often is understood as a candidate for full automation. Our design is instead based on the work practice of human analysts, and intended to afford a human analyst tools to work with the text smoothly and painlessly, leaving the human effort to be expended on the most crucial and demanding task of content analysis, but freeing the analyst from keeping track of consistency.

Design principle 2: Incremental refinement in clustering pipeline. The assumption of interaction designers is often that users are best served by automation. Our design is a departure from that assumption. We want our system to go beyond a one-shot dialog. The dialog builds on incremental specialization of the analysis: in a few iterations of the data set, the analyst can achieve a stable clustering to save and report.

Design principle 3: Errors do not matter. The assumptions made by the system, however well its algorithms are designed and however well established its background knowledge is, are often daring and sometimes mistaken. The design is intended to display analyses, and to allow the analyst to correct misclusterings with little effort, with a high degree of interactivity. The above principle of incremental refinement alleviates the presence of errors — the analyst is able to find themes in the texts, even if some of the first clusters were irrelevant or overlapping.

Design principle 4: Representation in surface terms. The end result of the analysis is a knowledge representation through which the set of texts can be understood better. This structure can be saved for future incoming data sets, e.g. a before-and-after study or a periodically repeated survey over some population. We want the knowledge representation to be inspectable, reportable, and editable by a human analyst without specialist knowledge. The representation is entirely in surface terms, for that purpose.

Design principle 5: No dependence on outside resources. We want the system to be portable to various languages, various domains of application, and various cultural areas. We do not want it to rely on costly or cumbersome lexical or encyclopædic resources which may not be available in all languages. The system is designed not to need anything but the texts under consideration and a larger sample of other background text written in the target language to tune term statistics.

IMPLEMENTATION

The functionality on which the system is built automatically clusters the documents into bins by lexical statistics. This creates clusters of documents that share topically important terms.

Text clustering. Lexical clustering builds on measures of term specificity to select which terms to use as clustering features, which requires general language data to be able to assess how specific or general a term is. Clustering by terms is fairly sensitive to genre-specific and topical usage, since a term which has high specificity in general language may have little utility in the context being examined.

Most standard lexically based clustering algorithms give similar results; we use a clustering algorithm based on insights from our previous research results on distributional semantics, [4] and we find that improving response speed and capacity of the system are more important to address (given Design principles 2 and 3 above) than marginal improvements in cluster quality.[5]

The example sentences from a hotel review data set given in (1) were all in the first iteration clustered together under the label *hotel*. A term such as *hotel* in hotel reviews does not appear to be a useful clustering feature. The texts should in most scenarios not end up being clustered on *hotel* but instead on *location* (for samples (1-a) and (1-c)) and *work* (for sample (1-b)) instead. Achieving this requires automatically reweighting term specificity during the clustering process, and, most importantly, as our system currently does, consulting the analyst to see if the clustering terms are appropriate and informative.

Manipulating clusters. Following the above design principles, the clusters are then displayed to the analyst for consideration. The main actions for the analyst are (1) joining existing closely related clusters, (2) discarding clusters that are of no interest, and

[4] The survey was performed in the Fall of 2016 to explore the attitudes to digital tools in teaching among students. http://www.berattarministeriet.se/undersokning/

[5] This is in keeping with earlier results comparing different text clustering systems, comparing their output with human assessments. There are differences, but they are comparatively small. [11]

(3) working on what terms characterise a cluster by approving synonyms suggested by the system or entering them manually.

The action of *joining* clusters into one common theme is a frequent operation to refine the end result, and our tool supports joining through simple direct manipulation. Similarly, clusters of low utility can be *discarded*, and the items constituting it are redistributed over other clusters instead. In this way, the content of the clusters are iteratively refined with simple and reversible point-and-click manipulation.

Synonyms. The nature of human language being as it is, we can expect many answers to diverge from the expected terminology. There will be many ways to say the same thing but you want them all in the same theme bin after the analysis process. The theme bin is represented by a set of terms which are prevalent in the texts clustered into that bin, and using a lexicon learned from text in the target language, [12] the system suggests synonyms to increase the coverage of that theme such as *friendly* for *pleasant* and related terms in a broad sense such as giving *coffee* and *pastry* for *breakfast* which will be of use in the example given in Figure (3). The analyst is also able to freely enter terms to enrich the representation of a theme.

(3)
- a. The staff were very *friendly* and *helpful*.
- b. The staff was *courteous* and *professional*, and they gave the impression that *hospitality* was something they enjoyed expressing.
- c. The staff was *personable* and demonstrated a true thankfulness for your business.
- d. The *breakfast* was always fine and we enjoyed a light *breakfast* every morning of a bowl of fruit together with a choice of a *bagel*, *toast* or *croissant*.
- e. However the hotel did offer free pastries, muffins, fruit, *coffee*, and juices every morning.
- f. There was no restaurant when we were there but they did offer *coffee* and *pastry* in the AM.

Multi-word terms. Most written languages build on white-space separated words, which is very convenient for tokenisation of the input stream in text processing. Many languages — and English is especially liberal in this respect — formulate multi-word compound terms quite freely, and all languages have set phrases such as *kick the bucket* and some degree of lexicalised multi-word terms, not least names such as *San Francisco* but also technical terms such as *linear accelerator* or *bed linen*. Our tool picks n-grams incrementally [12], as they appear in streaming data, and uses this to propose multi-word terms found in the text.

Handling several languages. Analysis of responses must as a rule be done in the language the responses were submitted. Our tool is built to be language agnostic and handles any human language (the only bottleneck being the quality of the synonym suggestions: to deliver reliable high-quality synonyms the system needs to have had access to some collection of general texts in the source language, such as a collection of newsprint, or a Wikipedia snapshot), and it still requires the analyst to be handy in the source language of the texts.

CASE STUDIES

We present here short abstracts of case studies where our tool as described above has been used. They serve to illustrate its versatility in application to multi-lingual and multi-cultural data, very open questions of wide-ranging themes, and drilling down into subthemes of customer reviews.

Attitude towards gender equality in seven cultural areas. In 2016, Gavagai was commissioned to execute a study in the Middle East, Latin America, Russia, and Sweden as part of an effort to monitor awareness of some aspects of Swedish society and Swedish foreign policy. The study collected 9800 free-text answers to open-ended survey questions in the various cultural areas and gave very various answers to questions such as the one given, with some sample answers, in (4). [5]

As one example we found a clear difference across cultural areas with respect to ”feminism”. The question as given in (4) gave very various attitudinal results. Explaining them by exploring the answers we found that feminism was associated with negative gender behavioural patterns such as machismo or with reverse discrimination in Latin American countries and in Russia, whereas it was accepted as a label for progressive policies and viewed comparatively positively in Middle Eastern countries. This analysis was made possible by identifying topical themes among the items with attitudinal loading.

(4) If a man or woman describes themselves as feminist, what would you think of that person? What kind of associations do you get? Is feminism positive or negative in your view? How would you describe feminism?
- a. *“Feminism is a positive concept, as women previously were discriminated against (earlier the world was sexist) whereas now women also find positions in areas which earlier were considered to be only for men.”*
- b. *“Feminism is neutral until it has acquired a mass character.”*
- c. *“I have a neutral view on this topic as each individual has their own perspective, as for me feminism shouldn’t exist in today’s world and education system.”*
- d. *“I consider feminism to be negative that it is the opposite to machismo or am I wrong?”*

“What do you most wish for the coming year?”. In order to better understand their customers’ thoughts and wishes for the coming year one of our customers, AMF – a limited liability life insurance company, – sent out a survey to more than 100,000 senior citizens with 14,793 responses.[2] The survey included the open-ended question:

(5) What do you most wish for the coming year?

Two thirds of the senior citizens responding to the open-ended question wished for a better health for themselves, followed by concerns about their family, the global society and peace. The hopes were expressed using a manifold of formulations as might be expected from a broad sample of senior citizens from all walks of life. Clustering those into consistent themes would be a major challenge for any human operator, but with the terminology support we found

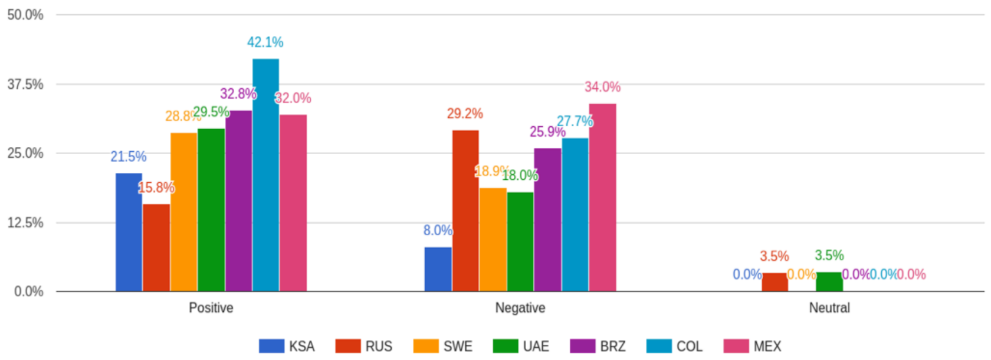

Figure 1: Attitude towards ”feminism” in seven cultural areas.

handily that there were strong underlying topics in the content. Of the top ten themes expressed, three or even four concern various aspects of money and economy.

”What makes airline passengers happy?”. We used our tool to analyse online consumer reviews of airlines published on an online consumer review site. [3] We collected 20 000 free text reviews of 22 airlines, with no quantitative data attached to them from the site. Attitude and topical themes are automatically identified and clustered. We measured how strongly opinionated reviewers are with regard to different aspects of their experience and we make these values comparable between different carriers. Some themes emerge from the text, with various degrees of prevalence for different airlines: Food, Drink, Seat, Service, Value, Inflight Entertainment, and so on and forth. Our main finding was that airline passengers seem to put up with almost anything, as long as they feel that they are being seen and looked after as individuals: the happiest passengers complained mostly about meals; the unhappiest about service. Satisfaction with staff service was a key driver for satisfaction with other aspects, such as the comfortability of the seat, the taste of the food, and for the overall passenger experience. This makes the results of review analyses much more actionable and shortens the path from attitude analysis to strategic business decisions.

LESSONS LEARNT

The advantages of using highly interactive automation for analysis is scale, speed, consistency, and saving human effort for the most important tasks.

An automated system has *full coverage*: all the above case studies would have been possible to do manually, if one single analyst or a very highly coordinated group of analysts had perused all the answers. This is impracticable at scale, unless automation is used.

An automated system is *quick*: an analyst is able to analyze more answers in a given time period which means that the number of responses to a survey can be larger which improves explanatory power. The granularity of the analysis can be increased to allow the analyst do drill down into more detailed and more actionable subtopics than before. The marginal effort for a larger survey increases sublinearly.

An automated system is *consistent*: it will allow one analyst to process more data, and provides the same initial results regardless of who does the analysis, reducing inter-rater variation. If a coding scheme is retained for repeated use, e.g. in monthly surveys, the analysis will remain consistent over time.

These factors reduce cost and increase reliability. Most important, however, is relieving the human analyst from the frustrating aspects of the coding task, freeing human effort to the more central task of understanding themes.

REFERENCES

[1] Douglass R Cutting, David R Karger, Jan O Pedersen, and John W Tukey. 1992. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

[2] Gavagai. 2016. *What do you most wish for the coming year?* Stockholm. http://gavagai.se/wp-content/uploads/2016/03/AMFPension-CustomerCase.pdf

[3] Gavagai. 2017. *What makes airline passengers happy?* Stockholm. http://gavagai.se/blog/2017/04/24/what-makes-airline-passengers-happy/

[4] Amaru Cuba Gyllensten and Magnus Sahlgren. 2015. Navigating the Semantic Horizon using Relative Neighborhood Graphs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

[5] Svenska institutet. 2016. *Feministisk utrikespolitik: rött skynke eller vit flagg?* Stockholm. https://si.se/wp-content/uploads/2016/12/Sverigebilden-Rapport-_om_synen_pa_-jamstalldhet.pdf (In Swedish; A slide deck with a summary in English is at http://gavagai.se/Gender_Equality_Study.pdf).

[6] Nick Jardine and Cornelis Joost van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information storage and retrieval* 7, 5 (1971).

[7] Sofus A. Macskassy, Arunava Banerjee, Brian D. Davison, and Haym Hirsh. 1998. Human Performance on Clustering Web Pages: A Preliminary Study. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD).*

[8] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 1999. A machine learning approach to building domain-specific search engines. In *Proceedings of the International Joint Conference on Artificial Intelligence*. IJCAI.

[9] Alicia O'Cathain and Kate J Thomas. 2004. ” Any other comments?” Open questions on questionnaires–a bane or a bonus to research? *BMC medical research methodology* 4, 1 (2004).

[10] Peter Pirolli, Patricia Schank, Marti Hearst, and Christine Diehl. 1996. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM.

[11] Dmitri G Roussinov and Hsinchun Chen. 1999. Document clustering for electronic meetings: an experimental comparison of two techniques. *Decision Support Systems* 27, 1 (1999).

[12] Magnus Sahlgren, Amaru Cuba Gyllensten, Fredrik Espinoza, Ola Hamfors, Jussi Karlgren, Fredrik Olsson, Per Persson, Akshay Viswanathan, and Anders Holst. 2016. The Gavagai Living Lexicon. In *Language Resources and Evaluation Conference*. ELRA.

[13] Mark Sanderson and Bruce Croft. 1999. Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

[14] Greg Schohn and David Cohn. 2000. Less is More: Active Learning with Support Vector Machines. In *Proceedings of the International Conference on Machine Learning*. ACM.