Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation

Magnus Rosell[*], **Sumithra Velupillai**[†]

[*] KTH CSC
100 44 Stockholm
Sweden
rosell@csc.kth.se
[†] DSV, KTH - Stockholm University
Forum 100
164 40 Kista
Sweden
sumithra@dsv.su.se

Abstract

Open answers in questionnaires contain valuable information that is very time-consuming to analyze manually. We present a method for hypothesis generation from questionnaires based on text clustering. Text clustering is used interactively on the open answers, and the user can explore the cluster contents. The exploration is guided by automatic evaluation of the clusters against a closed answer regarded as a categorization. This simplifies the process of selecting interesting clusters. The user formulates a hypothesis from the relation between the cluster content and the closed answer categorization. We have applied our method on an open answer regarding occupation compared to a closed answer on smoking habits. With no prior knowledge of smoking habits in different occupation groups we have generated the hypothesis that farmers smoke less than the average. The hypothesis is supported by several separate surveys. Closed answers are easy to analyze automatically but are restricted and may miss valuable aspects. Open answers, on the other hand, fully capture the dynamics and diversity of possible outcomes. With our method the process of analyzing open answers becomes feasible.

1. Introduction

Questionnaires are an important source for new research findings in many scientific disciplines, as well as for commercial exploitation. They may contain both closed ended and open ended questions. The answers to these are called closed and open answers, respectively. Closed answers are restricted to a fixed set of replies, while open answers are not. Statistical methods can be used to study closed answers in large questionnaires. Open answers must be reviewed manually.

Open answers contain valuable and detailed information that is very time-consuming to analyze manually. Methods for assisting the process of analyzing open answers in questionnaires are needed. Natural Language Processing tools could aid such processes, by enhancing the quality of the methods and therefore also the end results.

In Text Mining methods for discovering new, previously unknown information from large text sets are studied (Hearst, 1999). One such method is text clustering, which divides a set of texts into groups (clusters) of texts with similar content. As the content of clusters usually is divers, human investigation and interpretation is needed. The investigation can be aided by the clustering method in several ways. For clustering to be really useful both textual and visual presentation of the clusters should allow the user to explore the results, and interactively focus on interesting and intricate parts.

Collecting large sets of demographic and lifestyle data systematically is central for epidemiological studies. In (Ekman et al., 2006) the feasibility of using web-based questionnaires is discussed. Moving towards e-epidemiology increases the possibilities of conducting large population-based studies immensely, both with respect to cost-efficiency and availability (Ekman and Litton, 2007).

We present a method for hypothesis generation using text clustering, involving human judgement in crucial steps. The method is applied to a large epidemiological questionnaire with promising results.

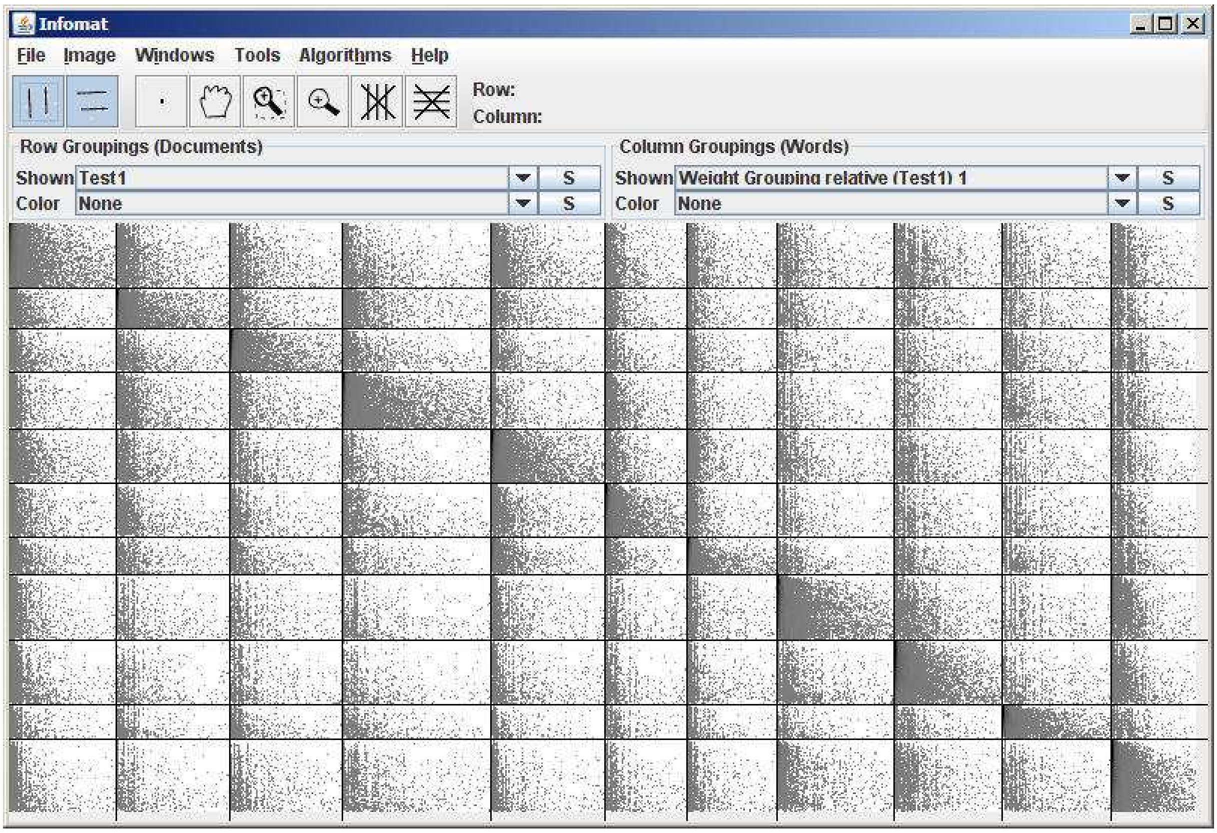
2. Related Work

Swanson and Smalheiser (1997) describe a method for hypothesis generation by linking possibly related medical literature. Their method exploits existing literature in order to discover previously unknown information and involves user interaction.

In the Scatter/Gather-system (Cutting et al., 1992) clustering is used as a tool for exploration of text sets. Clusterings are presented in a textual format and the user can interactively choose to re-cluster parts of the result, homing in on interesting themes.

To our knowledge, little research has been performed on automatically revealing new information from open answers in questionnaire data. Li and Yamanishi (2001) present a method for analyzing open answers in questionnaires using rule analysis and correspondence analysis. They describe a few other systems, but information about these is not readily found.

Central to all exploration methods is human interaction. Exploration of unstructured information requiers human interpretation.



Row Groupings (Documents)
Shown Test1
Color None

Column Groupings (Words)
Shown Weight Grouping relative (Test1) 1
Color None

Figure 1: Infomat. 41 549 texts (rows) from the questionnaire presented in Section 4. clustered to 11 clusters (K-Means), represented by 5 978 words (columns). Clusters are separated by lines. The text clusters are sorted according to smoking purity, where those with the highest amount of smokers are found at the top. The texts in each cluster are sorted in order of similarity to the cluster centroid. The words are clustered using the algorithm of Figure 3. Within each word cluster the words are sorted in order of weight in the corresponding text cluster centroid. A distinct diagonal is visible in the 11-by-11 pattern as could be expected. (The opacity of each pixel is proportional to the sum of the weights of its matrix elements.)

3. Method

We propose a method for hypothesis generation from open answers in questionnaires based on text clustering. The method could be described as follows:

1. Cluster the text set
2. Identify interesting clusters
3. Explore cluster contents
4. Formulate potential hypotheses

These steps should be repeated several times. For each repetition different settings (text representation, different clustering algorithms, etc.) could be used. Any recurring hypotheses may be further studied, through literature studies or new surveys.

The proposed method is semi-automatic and can easily be applied using the Infomat tool (see Section 3.1.). User interaction is a central part of the process. Human judgement is required to draw relevant conclusions in each step.

3.1. The Infomat Tool

Infomat[1] is a vector space visualization tool aimed at Information Retrieval (IR) and text clustering in particular (Rosell, 2007). It incorporates the ideas from the Scatter/Gather-system (Cutting et al., 1992), adding new functionality.

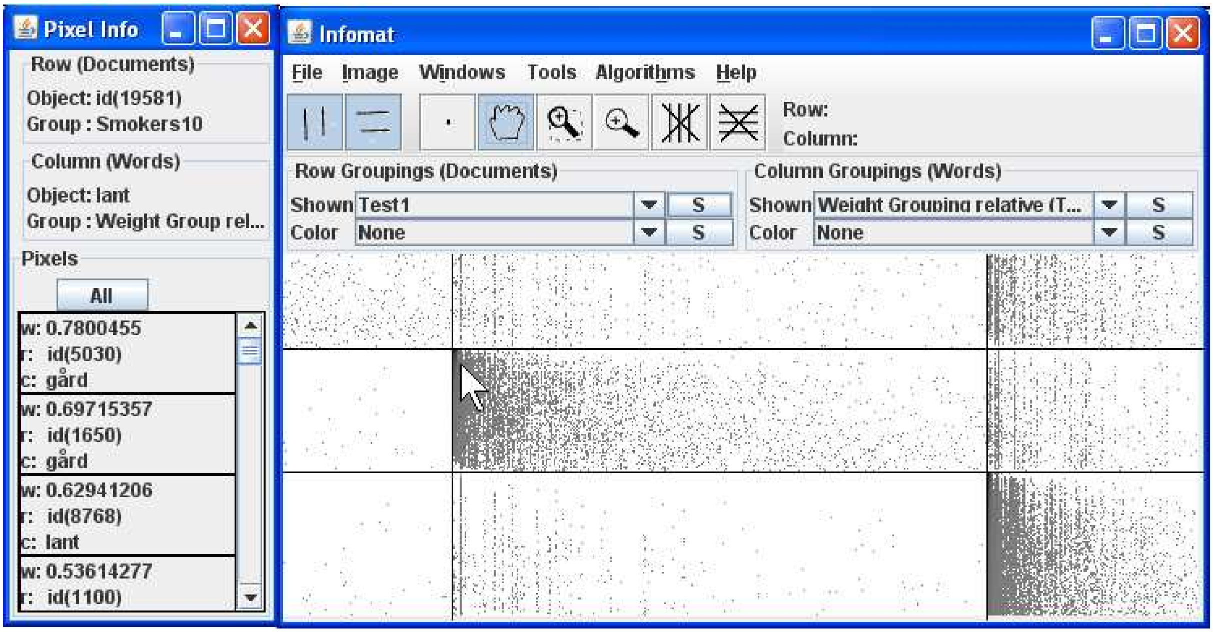
Infomat presents information stored in a matrix as a scatter plot, where the opacity of each pixel is proportional to the weight(s) of the corresponding matrix element(s). Here texts are represented in the vector space model by a text-by-word matrix, see Figure 1 for an example.

By sorting the rows (texts) and columns (words) in different ways hidden relationships between the objects may be exposed as visual patterns. Since the rows and columns represent actual objects (texts and words), the visual patterns are possible to comprehend.

Textual information about the matrix can be obtained in different ways. For instance the text(s) and word(s) of each pixel are presented when the cursor is moved over the matrix. It is also possible to zoom in and out, in order to investigate parts of the matrix in more detail, see Figure 2.

Infomat allows the user to cluster both rows and columns. The algorithm introduced in Figure 3 constructs a clustering of the words relative to a text clustering. An extensive description of the content of a text cluster is given by the combination of the visual patterns and the corresponding relative word cluster. (Naturally, reading the actual texts in the clusters can provide further insights.)

[1] http://www.csc.kth.se/tcs/projects/infomat/infomat/



Row (Documents)
Object: id(19581)
Group : Smokers10

Column (Words)
Object: lant
Group : Weight Group rel...

Pixels
All

w: 0.7800455
r: id(5030)
c: gård

w: 0.69715357
r: id(1650)
c: gård

w: 0.62941206
r: id(8768)
c: lant

w: 0.53614277
r: id(1100)

Row:
Column:

Row Groupings (Documents)
Shown Test1
Color None

Column Groupings (Words)
Shown Weight Grouping relative (T...
Color None

Figure 2: Infomat zoom example. A part of the picture in Figure 1 (centered around the second row and column clusters from the bottom right corner) is shown in the Infomat main window. The *Pixel Info* window to the left gives the matrix elements (weight, text, word) that are represented by the pixel indicated by the cursor. It also shows to which groups (along rows and columns) the texts and words belong. The Swedish word *gård* means “a farm” and *lant* could be translated to “country”. There are several more words in the scroll list.

```
Input:  a text set $\mathfrak{T}$,
        a set $\mathfrak{W}$ of all words appearing in $\mathfrak{T}$,
        a clustering of the texts $\{T_i\}$.

  • For each text cluster $T_i$:
      – calculate the centroid $\overline{T_i}$
      – construct an empty corresponding word cluster $W_i$
  • For each word $w \in \mathfrak{W}$:
      – find $\overline{T_k}$ with maximal weight for $w$
      – put $w$ in $W_k$, ordered by its weight in $\overline{T_k}$

Output:  a clustering of the words $\{W_i\}$.
```

Figure 3: *The Relative Clustering Algorithm*

3.2. Identifying Interesting Clusters

A closed answer in a questionnaire may be viewed as a categorization, making it possible to measure clusterings of open answers by ordinary clustering quality measures. If the categorization distribution (measured by a quality measure) in a cluster differs significantly from the entire set the cluster is potentially interesting. Whether a categorization distribution in a cluster differs sufficiently must be judged by the user and depends on the data set, the categorization, etc. In Infomat the clusters can be sorted in order of quality measure value, identifying the clusters with extreme values as the most interesting, see Figure 1 for an example.

In the context of clustering the quality measure *precision* (p) compares each cluster i to each category j in the categorization:

$$p_{ij} = \frac{n_{ij}}{n_i}, \qquad (1)$$

where n_{ij} is the number of texts from category j in cluster i, and n_i is the number of texts in cluster i. From the dominating category we get the *purity* for each cluster:

$$\rho_i = \max_j \{p_{ij}\}. \qquad (2)$$

The purity is a useful measure here as it is easy to understand. This helps in formulating the hypothesis, see Section 3.4..

3.3. Exploring Cluster Contents

One of the main challenges in text clustering is to describe the contents of the clusters to a user. Other text clustering tools, Scatter/Gather (Cutting et al., 1992) for instance, usually only present a headline consisting of some of the words with the highest weights in the cluster. However, short cluster headlines only provide a partial description of the cluster content, possibly omitting important characteristics.

For each text cluster a corresponding relative word cluster created by the algorithm in Figure 3 constitutes a cluster description. It provides an extensive overview of the cluster content, which can be grasped through browsing with the Infomat tool, as described in Section 3.1..

3.4. Formulating Hypotheses

If a cluster is deemed interesting, as described in Section 3.2., a hypothesis can be formulated from the cluster content (Section 3.3.). It can be expressed as a relation between the content and the closed answer distribution in the cluster. A hypothesis that recurs over several method iterations is worth investigating further.

3.5. Filtering Hypotheses

The generated hypotheses should be seen as starting points for further analysis. Therefore the exact quality measure values (in the identification of interesting clusters) are not that important – it is the tendencies that matters. Further, the hypotheses might not be novel as they are constructed solely from the investigated questionnaire. A domain expert can make well judged decisions on which tendencies to further pursue.

If the method produces an interesting hypothesis it can be considered useful. Whether the hypotheses holds can only be determined through further studies on material separated from the questionnaire.

3.6. Method Extensions

The method could be extended in several ways. In fact, the more ways the data is processed (revealing the same relations) the better. Several clusterings of rows and columns using different clustering algorithms can provide insights when combined. An especially interesting clustering technique, which is clearly related to the relative clustering algorithm in Figure 3, is *co-clustering* (Dhillon, 2001), where text and word clusterings are constructed simultaneously.

In the identification of interesting clusters, other quality measures, for instance *entropy*, could be used. They could be interesting as an aid in a general investigation of the text set. It is, however, harder to formulate a hypothesis using more abstract and complex measures than purity.

Several closed answers could be used in the identification of interesting clusters, for instance by constructing a categorization of the combination of them. If several open answers are available, clusterings of them could be used as well, considering any one of them a categorization. Further, the Infomat tool allows the user to view a second clustering or categorization along both rows and columns by coloring matrix elements depending on which cluster/category they belong to.

As presented here, the method relies heavily on human judgement. We believe this is unavoidable (and even desirable). Still, perhaps a more automated process could aid the human further in making these judgements. For instance, a predefined scheme of clusterings (and re-clusterings of parts of clustering results) could be run. The results of these could be presented in a condensed form, by for instance only displaying clusters that have been deemed sufficiently interesting automatically. This would make the identification of recurring relations more straightforward.

4. Text Set: Questionnaire

Karolinska Institutet (Swedish Medical University) administrates The Swedish Twin Registry[2], the largest twin registry in the world, containing information about more than 140 000 twins. See (Lichtenstein et al., 2002; Lichtenstein et al., 2006) for a description of the contents and some findings that have come from it.

	Gender	Smoking
ρ	0.52 (women)	0.71 (non-smokers)

Table 1: Gender and smoking purity for the entire set

	Women	Men
ρ	0.75 (non-smokers)	0.65 (non-smokers)

Table 2: The purity of smokers by gender for the entire set

The registry is based on information from questionnaires containing both closed and open answers. The combination of these provides a large set of valuable (medical, biological, sociological, etc.) information. Manual treatment of this is slow and costly.

The work presented here does not focus on revealing twin-specific information. Instead, the text set is used as an example to show how questionnaire data can be exploited.

4.1. An Open Answer on Occupation

Between 1998 and 2002, all twins born in or before 1958 were asked, among other things, to describe their occupation in a few words or sentences (in Swedish). The described occupation is either the last or the primary occupation during the respondent's lifetime. These answers provide a large set of texts with valuable but unaccessible information.

4.2. Representation of the Open Answer

In our experiments we have used the vector space model with tf*idf-weighting to represent the texts and the cosine measure for calculating similarity between texts and clusters. After applying a stoplist, we split compounds using the spell checking program STAVA (Kann et al., 2001) and conduct lemmatization using the grammar checking program Granska[3]. In (Rosell, 2003) improvements in clustering results on Swedish news texts using such techniques are reported.

After preprocessing 41 549 texts remained, having on average 10 different words (including compound parts). There were only 5 978 different words in total and each word occurred in on average 69 texts[4].

4.3. Closed Answers: Gender and Smoking

The questionnaire has several closed answers regarding smoking habits. We have constructed a categorization where we define *smokers* as respondents that have smoked more than a year, and *non-smokers* as all other. There are 12 244 smokers, that is 71% are non-smokers. Table 1 gives the smoking and gender purity for the entire set, and in Table 2 the purity of smokers by gender is shown.

[2] http://www.meb.ki.se/twinreg/index_en.html

[3] http://www.nada.kth.se/theory/projects/granska/

[4] After removing words that only occur in one text.

Clusters	Cluster A	Cluster B	Cluster C	Cluster D
Words	boss (chef) leader (ledare) personell (personal) company (företag) work- (arbets) task (uppgift) administrative (administrativ) lead (leda) project (projekt) responsibility (ansvar)	drive (köra) chauffeur (chaufför) car (bil) driver (förare) lorry- (lastbils) lorry (lastbil) truck (truck) taxi (taxi) load (lasta) road carrier (åkeri)	assistant (biträde) care (vård) home (hem) food (mat) old (gammal) cook (laga) help (hjälpa) service (tjänst) sick (sjuk) housing (boende)	country (lant) forest (skog) farm (gård) cultivator (brukare) animal (djur) agriculture (lantbruk) agriculture (jordbruk) cow (ko) worker (arbetare) works (bruk)
Number of texts	3747	2037	4083	2231
Number of words	3358	2483	2706	2137
ρ(non-smokers)	0.64	0.65	0.76	0.78
ρ(gender)	0.73 (men)	0.90 (men)	0.91 (women)	0.64 (men)

Table 3: Example text clusters from a clustering to 20 clusters of the occupation answers. The two top and two bottom clusters sorted in order of smoking purity. The words are the highest ranked in the corresponding word clusters and have been manually translated from Swedish. The sizes of the text and relative word clusters, as well as the smoking and gender purity are also displayed.

5. Experiment

We have applied our method on the questionnaire, described in the previous section, using the Infomat tool with the K-Means algorithm. The latter since it is fast, which makes the waiting times quite acceptable and the exploration pleasant even on an ordinary home computer.

We clustered the open answers regarding occupation several times to different numbers of clusters. Each time we also applied the relative clustering algorithm (see Figure 3) to the words. An example clustering is given in Figure 1. We also compared each clustering to the closed answer to identify interesting clusters as described in Section 3.2. The text clusters of Figure 1 are sorted in order of purity of smokers – the higher up in the picture the more smokers in the cluster.

We browsed the cluster contents as described in Section 3.1. In this particular example the cluster second from the bottom caught our eye: it has a low percentage of smokers, it is small and seemed to be coherent. In Figure 2 we have zoomed in on this cluster (and its relative cluster). After further browsing at this level we became convinced that a substantial part of the answers described occupations related to farming. Hence, we formulated a potential hypothesis, a relation between the open and closed answer: farmers smoke less than the average.

We repeated the steps of our method several times and observed the same relation in many of the iterations. Table 3 gives a textual presentation of another clustering, where *Cluster D* further supports this discovery.

After only a few hours[5] of exploration, concentrating on the most interesting clusters, we have formulated the following four hypotheses. They correspond well to the four clusters presented in textual form in Table 3.

A People working in leadership positions smoke more than the average.

B People working in the transportation industry smoke more than the average.

C Care workers smoke less than the average.

D Farmers smoke less than the average.

In the next section we try to assess hypothesis D, which was most consistent. The others may in part be explained by the gender distribution, see Tables 2 and 3, and should be studied further.

Studying the text clusters in Table 3, compared to gender regarded as a categorization, four other hypotheses could be formulated. We leave it to the reader to assess the quality of these.

6. Evaluation

With no prior knowledge of smoking habits in different occupation groups we have generated a hypothesis indicating a tendency that farmers smoke less than the average. In order to support or discard it thorough investigations and/or surveys should be performed. Lacking such possibilities, we have tried to find existing comparable surveys on smoking habits (after formulating the hypothesis).

Surveys differ in what they cover, both population sample and questionnaire formulation. The definition of a *smoker* may vary between surveys. Also, there exist many categorization systems for occupations, many of them differing in specificity and structure.

The questionnaire we have derived our hypothesis from is described in Section 4.. We have found the following comparable surveys:

- a Swedish survey by Statistics Sweden (SCB, 2006)
- two U.S.A. surveys (Lee et al., 2004; Lee et al., 2007)
- a European survey (McCurdy et al., 2003)
- an Australian survey (Smith and Leggat, 2007)

The most comparable survey is the one made by Statistics Sweden[6] (SCB), as it is conducted on the Swedish population. SCB is the central government authority for official statistics in Sweden. They provide general population statistics.

[5] Naturally, the amount of time can differ significantly depending on the questionnaire and the purpose of the investigation. The experiment demonstrates that interesting results can be obtained within a reasonable time.

[6] http://www.scb.se

The survey performed by SCB covers the years 1980 – 2005 and the ages 16 – 84. It is given almost every year and the statistics are presented from different aspects: household type, age groups, socio-economic group, etc. Here, smokers are defined as respondents who smoke daily. We focus on the years 1998 – 2003 (the time for the twin questionnaire) and the statistics for farmers as a socio-economic group.

The percentage of smokers overall in the SCB survey is smaller than in the questionnaire, as well as among farmers, see Table 4. However, the tendency that farmers smoke less than the average can also be seen here. Thus, the SCB survey supports our hypothesis.

	All workers	Farmers
SCB 1998 – 99	23.9%	8.7%
SCB 2000 – 01	24.6%	7.2%
SCB 2002 – 03	23.4%	8.9%
Questionnaire	29%	-
Cluster D	-	22%

Table 4: SCB: daily smokers in socio-economic groups in Sweden 1998 – 2003, ages 16 – 84. Questionnaire: smokers (according to definition in Section 4.3.) among twin respondents 1998 – 2002, born in or before 1958. Cluster D: one cluster from a clustering of the open answers in the questionnaire, see Table 3.

All surveys have different occupation categorization systems. The U.S.A. surveys, for instance, utilize a fine-grained categorization of farmers, and the portion of smokers differs between the subgroups. Also, the surveys cover different age groups. The European survey is focused on a younger population sample. Further, different smoker definitions are used. The Australian survey distinguishes *current*, *ex-*, and *never*-smoker groups. However, the tendency that farmers smoke less than the average is apparent in all surveys.

Considering all differences between the surveys and the twin questionnaire we can confirm our hypothesis, that farmers smoke less than the average. Thus our method is proven successful.

7. Conclusions and Future Work

We have presented a method for hypothesis generation from questionnaires through text clustering evaluation. Using the method we have generated the hypothesis that farmers smoke less than the average, which we have confirmed through literature studies. Normally, a new investigation would need to be performed.

We plan to apply the method on other questionnaires in different domains. Also, it could be applied on other types of data sets containing both textual data and data restricted to predefined values. One interesting example is electronic medical records.

Our method makes it feasible to explore and analyze open answers in large questionnaires, potentially containing hidden information. It provides a means for interactively revealing interesting parts of that information, reducing the manual work load significantly.

Acknowledgements

Our thanks go to VR (Vetenskapsrådet - the Swedish Research Council) for funding and to all participants in the project “Infomat” from KTH CSC (The School of Computer Science and Communication at The Royal Institute of Technology) and the department of Medical Epidemiology and Biostatistics (MEB) at Karolinska Institutet, both in Stockholm, Sweden.

We also thank VINNOVA (Swedish Governmental Agency for Innovation Systems) for funding and all members of the project KEA (Knowledge Extraction Agent) from DSV, KTH-Stockholm University and KTH CSC.

Furthermore, we would like to thank GSLT (The Swedish Graduate School of Language Technology).

The Swedish Twin Registry is supported by grants from VR, the Department of Higher Education, and NIH (grant AG-08724).

8. References

D. R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey. 1992. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proc. 15th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*.

I. S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD ’01: Proc. 7th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pages 269–274, New York, NY, USA. ACM.

A. Ekman and J. E. Litton. 2007. New times, new needs; e-epidemiology. *European Journal of Epidemiology*, 22:285–292(8).

A. Ekman, P. Dickman, Å. Klint, E. Weiderpass, and J. E. Litton. 2006. Feasibility of using web-based questionnaires in large population-based epidemiological studies. *European Journal of Epidemiology*, 21:103–111(9).

M. A. Hearst. 1999. Untangling text data mining. In *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, pages 3–10, Morristown, NJ, USA. Association for Computational Linguistics.

V. Kann, R. Domeij, J. Hollman, and M. Tillenius, 2001. *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Ludek Hrebicek*, volume 60, chapter Implementation aspects and applications of a spelling correction algorithm.

D. Lee, W. LeBlanc, L. Fleming, O. Gómez-Marín, and T. Pitman. 2004. Trends in US smoking rates in occupational groups: the national health interview survey 1987-1994. *J. Occup. Environ Med.*, 46(6):538–48.

D. Lee, L. Fleming, K. Arheart, W. Leblanc, A. Caban, K. Chung-Bridges, S. Christ, K. McCollister, and T. Pitman. 2007. Smoking rate trends in U.S. occupational groups: The 1987 to 2004 national health interview survey. *J. Occup. Environ Med.*, 49(1):75–81.

H. Li and K. Yamanishi. 2001. Mining from open answers in questionnaire data. In *KDD ’01: Proc. 7th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pages 443–449, New York, NY, USA. ACM.

P. Lichtenstein, U. De faire, B. Floderus, M. Svartengren, P. Svedberg, and N. L. Pedersen. 2002. The swedish twin registry: a unique resource for clinical, epidemiological and genetic studies. *Journal of Internal Medicine*, 252:184–205.

P. Lichtenstein, P. F. Sullivan, S. Cnattingius, M. Gatz, S. Johansson, E. Carlstrom, C. Bjork, M. Svartengren, A. Wolk, L. Klareskog, U. de Faire M. Schalling, J. Palmgren, and N. L. Pedersen. 2006. The swedish twin registry in the third millennium: An update. *Twin Research and Human Genetics*, 9(6):875–882.

S. A. McCurdy, J. Sunyer, J. Zock, J. M. Antó, and M. Kogevinas. 2003. Smoking and occupation from the European community respiratory health survey. *J. Occup. Environ Med.*, 60(9):643–8.

M. Rosell. 2003. Improving clustering of Swedish newspaper articles using stemming and compound splitting. In *Proc. 14th Nordic Conf. on Comp. Ling. – NODALIDA '03*.

M. Rosell. 2007. Infomat – a vector space visualization tool. In M. Sahlgren and O. Knutsson, editors, *Proc. of the Workshop Semantic Content Acquisition and Representation (SCAR) 2007*. Swedish Institute of Computer Science (SICS), Stockholm, Sweden. SICS Technical Report T2007-06, ISSN 1100-3154.

Statistics Sweden SCB. 2006. Undersökningarna av levnadsförhållanden (Living condition survey). http://www.scb.se/LE0101.

D. Smith and P. Leggat. 2007. Tobacco smoking by occupation in Australia: Results from the 2004 to 2005 national health survey. *J. Occup. Environ Med.*, 49(4):437–445.

D. R. Swanson and N. R. Smalheiser. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Intell.*, 91(2):183–203.