

# The Image of Germany in the World

## An Email and Web Mining Approach

Bettina Berendt, Mark Draheim

**In this paper, we propose that email corpora can be as indicative of large-scale Internet user interests and concerns as social media are, and that by being less public than most wikis or blogs, they may also capture voices that are otherwise unheard. We analyse an email corpus from a major German news service whose official role it is to 'present Germany abroad', defining images that people could have of Germany. We propose that the emails sent in response to this reflect, to a certain extent, the image that people do have of Germany. Building on the email corpus and on Web site data, and using different forms of LSA-based data analysis, we identify salient topics. We develop measures of pattern interestingness that help to focus the analysis on novel, useful, valid and understandable patterns in the data. We suggest that these methods can also be employed in investigations of other media, producing a multi-perspective view of Internet users' concerns, thoughts and expectations.**

## 1 Introduction

During the last years, Internet users have found their voice in unprecedented numbers. Most visible has been the explosion in the number of social-media sites, wikis, and blogs. In parallel with this development of data, data analysis by data mining methods has become increasingly popular. However, this focus on new media of interaction neglects the self-selection bias inherent in these media, and it overlooks the continuing popularity of email and its significance for people who prefer addressing one (even if abstract) concrete communication partner, over addressing a group of people or even the world at large.

Unlike contributions written with the public in mind (such as a typical blog or wiki, or a letter to the editor of a newspaper or Web site), emails are often not written to express a firm opinion on something, to be shared with the world. Emails allow people to express questions, non-understanding, or requests. In this sense, they reflect expectations more than opinions, and they may indicate (potentially vague) associations one has with a core topic, rather than show the aspects of that core topic about which one has formed definite opinions. The application example of this paper, emails sent to a major German news service whose official role it is to 'present Germany abroad', captures opinions in this sense. This interpretation, gained by reading a sample of emails, led to the application question of this paper: How can data mining help to structure a large corpus of texts by the topics they discuss?

This is a typical question for text mining, which has been an active research area during the past years. Many powerful mining methods exist for classifying or grouping texts and for finding local patterns in them. In this paper, we focus on one group of such methods, latent semantic indexing / analysis (LSA), that are particularly suited for highly exploratory searches for structure in text corpora.

However, a typical LSA analysis will return many such topics, and as in other mining applications, the question arises which ones are really interesting. In other areas of data mining, measures of pattern interestingness have been proposed

to help the analyst and/or the method to focus on important patterns.

The contributions of this paper are threefold: First, we use two LSA tools for analysing the given email corpus. We argue that (certain) emails can serve as a valid source of information or 'buzz metric', comparable to more standard social media. Second, we propose new interestingness measures to interpret and profit from such analyses, and we illustrate how different tools support the assessment of these measures in different ways. By this, we aim to contribute to the discussion of suitable tools for analysing Internet users' concerns. Third, we do this guided by a case study that is interesting in its own right, as an indicator of 'the images people have of Germany'. After a short overview of related work, the analysis is presented in Section 3. The interestingness measures are defined and applied, and the tools evaluated, in Section 4. Section 5 concludes with an outlook.

## 2 Related work

Related work comes from text mining, email mining, and interestingness measures. For reasons of space, we cannot discuss the huge area of text mining here; instead, we give a short overview of the two methods that will be used in Section 3.

*Latent semantic analysis* ([10], for an introduction see [14]) is a statistical method that tries to find semantic relationships between terms in large text corpora by analysing co-occurrences of terms (words or compounds / n-grams). The aim is to find similar meanings regardless of the choice of words (polysemy, synonymy). LSA represents a corpus in a term-document matrix where each document is a vector in a high-dimensional term space. The dimensions are then reduced using singular value decomposition. Terms and documents can be projected into the reduced space. Terms that are projected onto the same dimension are assumed to be semantically related. Therefore, the reduced dimensions are also called topics or concepts. We use the term *topics* in the present article.