A new dimension for user modeling based on the use of sensory vocabulary

Gudrun Kellner[1], Bettina Berendt[2]

[1] Vienna University of Technology, Institute of Software Technology and Interactive Systems, E-Commerce Group, Favoritenstr. 9-11/188, 1040 Vienna, Austria kellner@ec.tuwien.ac.at
[2] K.U. Leuven, Department of Computer Science, Celestijnenlaan 200A, 3001 Leuven-Heverlee, Belgium bettina.berendt@cs.kuleuven.be

Abstract. In our research, we investigate the use of sensory vocabulary in forum texts as a source for implicit information on the user. Therefore, a corpus with more than 1,000,000 forum posts was analyzed for the occurrence of expressions that are directly linked to a sensory system. We found that users differ significantly in their use of sensory expressions and that most users have preferred patterns for the use of sensory expressions.
Keywords: User Modeling, Perception, Language use

1 Introduction

Digital media permits to present information in manifold ways. An adequate mode of information presentation helps the user with information processing, e.g. some users prefer to read new information, some prefer listening to it. Research shows that information can be more easily understood if its presentation is adapted to the cognitive style of the target user [2]. One aspect of cognitive preference consists in the preferred mode of perception [2;3]. Such information might be of interest for every user model used in a setting where the user's interest needs to be captured or the user's process of information perception and organization shall be supported [4].

But how can such a preference be elicited? We propose to use existing text, written by the target user and published in the internet, as a source of information about the user's perceptual preference. We follow the tradition of lexicon based approaches [6;8] such as used in opinion and sentiment mining when we look for expressions with a link to a sensory system, e.g., "green" or "see" as visual, and "loud" as auditory words.

2 Method

Our work proposes an extension to existing user models. We present a modeling technique that implicitly acquires sensory vocabulary data by analyzing forum text.

Sensory lexicon We based our lexicon of sensory vocabulary on the list of stems of sensory vocabulary collected by [5]. The list of sensory stems may be divided into 4 disjoint sets, namely the lexicon of visual vocabulary L_V, of auditory vocabulary L_A, of kinesthetic vocabulary L_K, and of olfactory and gustatory vocabulary $L_{\{OG\}}$.

Measures The use of sensory vocabulary in a document p is expressed as a four-dimensional vector *vprofile(p) = [p.V, p.A, p.K, p.OG]* based on the frequency of sensory vocabulary per sensory system. Thus, the visual component of the vector (the other components are calculated analogously) is defined as

$$p.V = \frac{|\{ t \in p' \mid v(t) = 1\}|}{|\{t \in p' | v(t) = 1 \text{ or } a(t) = 1 \text{ or } k(t) = 1 \text{ or } og(t) = 1\}|} \quad (1)$$

where *t* are the terms in *p'*, which is *p* modeled as a bag of words, and *v(t)* (etc.) are the sensory indicators of the term:

$$\text{indicator level } s(t) = \begin{cases} 1 & \text{if stem(t) is in the lexicon of sensory vocabulary} \\ 0 & \text{in all other cases.} \end{cases} \quad (2)$$

VAKOG profiles *vprofile(p)* were not only calculated for each document, but also for each author by concatenating all posts of this author to one new pseudo-document and calculating its profile as described in Equation (1). The similarity *vsim(p1,p2)* between two posts was measured as the cosine similarity of their profiles.

Hypothesis Every user has a preference profile for sensory modalities, expressed as a profile of usage of sensory vocabulary. Hence, the similarity *vsim(p1,p2)* of posts written by one author should be higher than to posts written by somebody else.

$$avg_{p1,\, p2 \in P,\, p1.author=p2.author}\ vsim(p1,p2) > avg_{p1,\, p2 \in P,\, p1.author \neq p2.author}\ vsim(p1,p2) \quad (3)$$

where the *pi* are posts, *P* is the set of all posts, *pi.author* is the post's author, and *vsim* is the similarity between the VAKOG profiles of its two arguments. We decided to control for content similarity by treating the full-text similarity of two posts as a covariate. The hypothesis then is refined to "if two pairs of posts each have the same full-text similarity, the pair of the same author will have higher VAKOG similarity than the unrelated pair". Full-text similarity was operationalized as the cosine similarity between the two posts modeled as bags of words (BOW) by the WEKA[1] StringToWordVector filter, weighted by TF.IDF [1].

Data To test the hypotheses, we chose Richling's forum corpus [7]. It is a corpus built on posts from discussion forums on the car type BMW E30, published in the years 2000 until 2007. This very narrow topic helps to minimize result variation due to discussion of different topics. The corpus is monolingual and consists of more than one billion posts in German, each post text accompanied by information on the author, the header, the reference post, and the date.

3 Results: The use of sensory vocabulary in forums

The E30 forum corpus consists of 1,053,841 posts, written by 30,021 different authors. A detailed distribution can be found in Table 1.

Concerning our hypothesis, we applied two methods of testing. (1) Comparison of distributions with Mann-Whitney's U test: Significance testing against the null hypothesis of equal distribution was calculated separately for BOW and for VAKOG similarity, comparing the hypothesis set with the set of all post pairs. (2) To combine the similarities BOW and VAKOG, we used loglinear modeling for 3-way contingency tables. The values of the three dimensions were: (i) pairwise full-text similarity, (ii) pairwise VAKOG similarity, and (iii) the 2 categorical values of the variable of interest (*author-* vs. *non-author-post-relation*). All results were

[1] http://www.cs.waikato.ac.nz/ml/weka

statistically significant with a p-value <.0001. The hypotheses testing on the E30 forum corpus shows a significantly higher VAKOG similarity within the hypotheses subsets than within the set of all post pairs. These results were consistent both for the comparison of means and the combined BOW and VAKOG similarities. This confirms our hypothesis and leads to the following conclusion: Authors of forum posts have a tendency to use sensory expressions in similar distributions over time. Hence, that distribution can be considered as an interesting extension to user descriptions for user modeling.

Table 1. The distribution of sensory expression in the E30 corpus.

Posts	1,053,841
Original posts	223,973
Answer posts	829,868
Not-empty posts (neP)	646,455
Av. nr of terms per neP	40.69
Authors	30,021
Av. nr of posts per author	35.10
E30 Dictionary*	
Different terms	474,264
Different visual terms	6,798
Different auditory terms	5,047
Different kinesthetic terms	7,518
Different olfactory+ gustatory terms	1,674
Sensory terms	785,303
Visual terms	318,305
Auditory terms	248,896
Kinesthetic terms	192,566
Olfactory+gustatory terms	25,536

*Terms are handled as different as soon as they differ in one letter, including typos versions

4 Conclusions and outlook

In our research, we have proposed a new dimension for user modeling based on the use of sensory expressions. Based on findings from cognitive information processing and learning styles, we investigated the potential of the idea to analyze the use of sensory expression as an individual preference that might indicate sensory preference.

We opted for an implicit approach to data acquisition concerning the use of sensory vocabulary by means of analyzing forum text. We found that authors tend to use sensory expressions in similar distributions when writing new posts.

The obtained results are quite encouraging: Our next steps are to enlarge the corpus of sensory expression, investigate the relation between preferred sensory system(s) and the use of sensory expression by combining forum text analysis with user tests on sensory preference, and examine the influence of topic on the use of sensory expression.

5 References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Harlow (1999)
2. Coffield, F., Moseley, D., Hall, E., Ecclestone, K.: Learning styles and pedagogy in post-16 learning. A systematic and critical review. LSRC, London (2004)
3. Dunn, R.: Commentary: Teaching Students Through Their Perceptual Strengths or Preferences. In: Journal of Reading 31/4, pp. 304–309 (1988)
4. Heinath, M., Dzaack, J., Wiesner, A., Urbas, L.: Applications for Cognitive User Modeling. In: Conati, C. et al. (Eds.): UM 2007. LNCS, vol. 4511, pp. 127–136. Springer, Heidelberg (2007)
5. Kellner, G.: Wege der Kommunikationsoptimierung. Anwendung von NLP im Bereich der Künstlichen Intelligenz. VDM, Saarbrücken (2010)
6. Nowson, S.: The Language of Weblogs: A study of genre and individual differences. Doctoral Thesis. University of Edinburgh (2006)
7. Richling, J.: Die Sprache in Foren und Newsgroups. VDM, Saarbrücken (2008)
8. Valitutti, A., Strappavara, C., Stock, O.: Developing Affective Lexical Resources. In: PsychNology 2/1, pp. 61–83 (2004)