17th European Conference on Artificial Intelligence (ECAI 06)
Brewka, Coradeschi, Perini, Traverso (Eds.)
pp. 815-816, ISBN 1-58603-642-4, ©IOS Press 2006.

Is Web Genre Identification Feasible?

Benno Stein[1] and **Sven Meyer zu Eissen**[1]

Abstract. This paper contributes to a facet from the area of Web Information Retrieval that has recently received much attention: The satisfaction of a user's personal information need with respect to text type, presentation type, or information quality. We imply that such properties can be quantified for all kinds of Web documents, and we subsume them under the term "Web genre" or "genre".

Recent surveys show that there is—to a certain degree—a common understanding of Web genre. However, the strictness by which genre and non-genre aspects of a document are experienced is an individual matter. To get a better understanding of the challenges of Web genre identification and its possible limits we investigate in this paper a very interesting question, which has not been posed by now:

Given a categorization $\mathcal{C}$ of documents (or bookmarks, links, document identifiers), can we provide a reliable assessment whether $\mathcal{C}$ is governed by topic or by genre considerations?

Keywords unsupervised learning, knowledge discovery, text mining, personal information retrieval

1 INTRODUCTION

Nearly all retrieval processes are topic-centered: We type in a keyword, provide a sample document, or browse a directory tree to get the desired piece of information. However, with the number of indexed documents develops the urgent need for information quality: Users are interested in certain *kinds* of information, or, as it is called here, in particular genres. In connection with text documents genre describes, among others, the set of conventions in the way in which information is presented, such as the style of writing, the presentation style, or the functional trait. An in-depth discussion of the term genre is beyond the scope of this paper,[2] but, for the time being it is sufficient to remember the following characterization: Topic and genre are orthogonal—or, with Dewdney [3]: "The form is the substance." This is illustrated in Figure 1, where the same articles of a newspaper page are classified under both topic and genre considerations.

Genre *identification* shall discover groups of texts that share a common form of transmission, purpose, or discourse properties [10, 12]. This means in the WWW context that genre identification can be understood as differentiation between research Web pages, personal experience reports, or commercial product information, for example. The application scenario of our paper connects at this point: Given a user's categorized document collection, $\mathcal{C}$ (in the form of bookmark folders for instance), we ask whether one is able to reliably determine the organization principle, say, the underlying categorization type behind $\mathcal{C}$: Is it topic or genre?

The remainder of this paper outlines Web genre research, provides technical background, and answers the question posed.

[1] Faculty of Media / Media Systems. Bauhaus University Weimar, Germany. benno.stein@medien.uni-weimar.de

[2] Santini has compiled an up-to-date discussion of this term [10].

Figure 1. The difference between topic and genre, illustrated at a newspaper page.

1.1 Related Work on Web Genre

There is little work on automatic Web genre identification, and, the question of feasibility is—if at all—answered indirectly only, following a simplistic three step approach: (*i*) definition of particular genre classes, (*ii*) compilation of a respective genre corpus, (*iii*) quantification of the learnability by constructing a classifier. In the following we organize the research in an ascending chronological order.

Bretan et al. propose a richer representation of retrieval results in Web search interfaces. Their approach combines content-based clustering and genre-based classification that employs simple part-of-speech information along with substantial text statistics. The features are processed with the C4.5 algorithm; the authors give no information about the achieved classification performance [2]. Based on an exploratory user study Roussinov et al. develop a genre scheme that comprises five genre classes: Topic, Publication, Product, Education, FAQ. Their work describes an ongoing study, and no discovery approach has been implemented [9]. Dimitrova et al. argue that shallow text classification techniques can be used to sort documents according to genre. The paper describes an ongoing study but experience related to classification performance is not reported [4]. Lee and Myaeng define seven genre types. Aside from Web-specific genres like Q&A or Homepage, the authors use also the newspaper-specific genres Reportage and Editorial. The feature set is a list of about hundred document terms tailored to the different genre classes [7]. Meyer zu Eissen and Stein report on a user study on Web genre usefulness from which they derive eight genre classes, which in turn form the building blocks of three genre profiles: Education, Geek, and Private. Within their comprehensive experiments classification performances between 60% and 80% were achieved [8]. Boese investigates the effects of Web document evolution on genre classification and poses the question: "How much do Web pages change over time within each genre?" The author answers this and related questions for two publicly available genre corpora [1].

Note that existing classification approaches treat not the problem that is addressed here; they start with the assumption that the type of the analyzed collection is a-priori known, namely, genre.

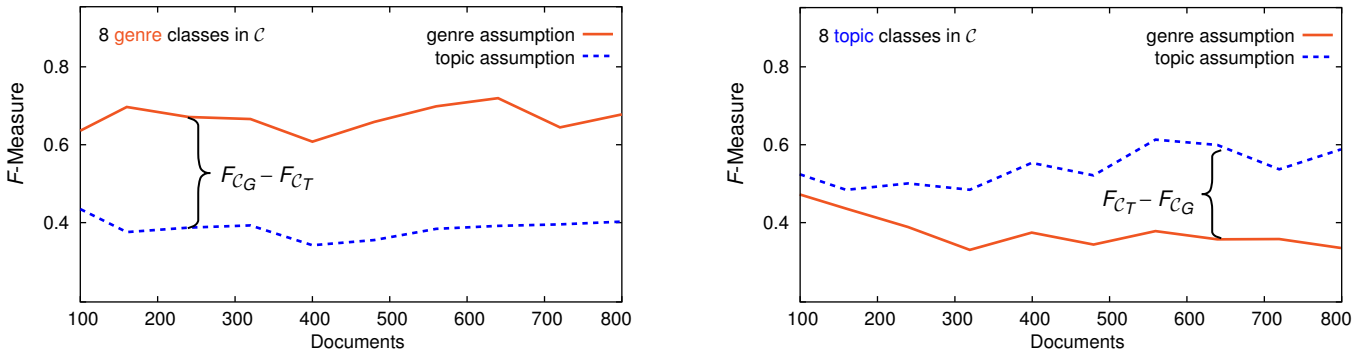

Figure 2. Plots that quantify the adequacy of the document models R_G and R_T. They unveil whether a categorization $\mathcal{C}$ is organized by genre or by topic.

2 DETERMINANTS OF A GENRE CLASSIFIER

With respect to the investigated features the existing approaches to genre identification fall into three groups: Classifiers that rely on a subset of a document's terms [11, 7], classifiers that employ linguistic features along with additional features related to text statistics and computational linguistics [6, 8], or both [5]. The following list gives an overview over the different feature types:

- customariness and style features
- part-of-speech and syntactic group analysis
- closed-class word sets and presentation-related features

Based on these features a powerful document retrieval model, R_G, for genre identification can be built. To keep the computational footprint of our genre model small we applied a discriminant analysis to select 18 features along with an appropriate weighting scheme.

By contrast, to capture the gist of a document with respect to its *topic*, the vector space model, R_T, is the most successful document retrieval model. It encodes a document d as a simple vector, which comprises weighted frequency values of the terms occurring in d.

3 HYPOTHESIZING CATEGORIZATION TYPES

Let D be a set of documents. An exclusive categorization $\mathcal{C} \subseteq \{C \mid C \subseteq D\}$ of D is a division of D into sets for which $\bigcup_{C_i \in \mathcal{C}} C_i = D$, and $\forall C_i, C_j \in \mathcal{C} : C_i \cap C_{j \neq i} = \emptyset$. The categorization $\mathcal{C}$ may be governed by topic or by genre considerations, and we introduce the following procedure to verify the underlying categorization type:

1. Construct for each $d \in D$ two document models, one under the genre document retrieval model, R_G, and one under the topic document retrieval model, R_T.
2. Based on a similarity measure (Euclidean or cos-similarity) construct two similarity graphs G_G and G_T. The edge weights in these graphs result from the similarity computations under R_G and R_T respectively.
3. Apply a clustering algorithm to the graphs G_G and G_T. The resulting clusterings are designated as $\mathcal{C}_G$ and $\mathcal{C}_T$.
4. Compute the F-measure (or another external reference measure) to quantify the congruence between $\mathcal{C}$ and $\mathcal{C}_G$ as well as between $\mathcal{C}$ and $\mathcal{C}_T$. The resulting values are designated as $F_{\mathcal{C}_G}$ and $F_{\mathcal{C}_T}$.
5. If $|F_{\mathcal{C}_G} - F_{\mathcal{C}_T}|$ is significant, $\mathcal{C}$ is organized under genre considerations if $F_{\mathcal{C}_G} > F_{\mathcal{C}_T}$, and under topic considerations otherwise.

The F-measure quantifies the degree of congruence between a (human) reference categorization $\mathcal{C} = \{C_1, \ldots, C_k\}$ and a clustering $\mathcal{C}' = \{C'_1, \ldots, C'_l\}$. The recall of cluster j with respect to category i, $rec(i,j)$, is defined as $|C'_j \cap C_i|/|C_i|$. The precision of cluster j with respect to category i, $prec(i,j)$, is defined as $|C'_j \cap C_i|/|C'_j|$. The F-measure of a clustering $\mathcal{C}'$, $F_{\mathcal{C}'}$, is:

$$F_{\mathcal{C}'} = \sum_{i=1}^{k} \frac{|C_i|}{|D|} \cdot \max_{j=1,\ldots,l} \{F_{i,j}\}, \text{ with } F_{i,j} = \frac{2 \cdot prec(i,j) \cdot rec(i,j)}{prec(i,j) + rec(i,j)}$$

A perfect clustering matches the given categories exactly and yields an F-measure value of 1.

Experiments Our experiments rely on the corpus of [8], where eight Web genre classes are distinguished: Help, Article, Discussion, Shop, Portrayal (priv and non-priv), Link Collection, and Download. The orthogonal topic categorization distinguishes the following eight topics: Sports, Annual results, International relations, Religion, Crime, Management moves, Money supply, Legal/judicial.

Based on this corpus we compiled 40 categorizations of different sizes and under both topic and genre considerations. It turned out that for each of these categorizations its type could be unambiguously determined by computing $|F_{\mathcal{C}_G} - F_{\mathcal{C}_T}|$, whereas a genre model comparable to [2] and as topic model the vector space model was employed. Figure 2 shows the developing of the respective F-measure values $F_{\mathcal{C}_G}$ and $F_{\mathcal{C}_T}$. The runtime complexity is dominated by the cluster analysis, and, using k-means, linear in $\mathcal{C}$.

Our current research focuses on document models for special retrieval situations and related learning strategies.

REFERENCES

[1] E. S. Boese and A. E. Howe, 'Effects of Web Document Evolution on Genre Classification', in *Proc. CIKM'05*, (2005).

[2] I. Bretan, J. Dewe, A. Hallberg, and N. Wolkert. Web-specific genre visualization, (1999).

[3] N. Dewdney, C. VanEss-Dykema, and R. MacMillan, 'The form is the substance: Classification of genres in text', in *Proc. ACL*, (2001).

[4] M. Dimitrova, A. Finn, N. Kushmerick, and B. Smyth, 'Web genre visualization', in *Proc. Human Factors in Comp. Systems*, (2002).

[5] A. Finn and N. Kushmerick, 'Learning to Classify Documents According to Genre', in *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, (2003).

[6] B. Kessler, G. Nunberg, and H. Schütze, 'Automatic detection of text genre', in *Proc. ACL and EACL*, Somerset, New Jersey, (1997).

[7] Y.-B. Lee and S. . Myaeng, 'Text genre classification with genre-revealing and subject-revealing features', in *Proc. SIGIR*, (2002).

[8] S. Meyer zu Eißen and B. Stein, 'Genre Classification of Web Pages: User Study and Feasibility Analysis', in *KI 2004: Advances in Artificial Intelligence*, volume 3228 LNAI, Springer, (2004).

[9] D. Roussinov, K. Crowston, M. Nilan, B. Kwasnik, J. Cai, and X. Liu, 'Genre based navigation on the web', in *Proc. 34th Hawaii International Conference on System Sciences*, (2001).

[10] M. Santini, 'State-of-the-Art on Automatic Genre Identification', Technical report, ITRI, University of Brighton, UK, (2004).

[11] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, 'Text genre detection using common word frequencies', in *Proc. Conf. on Computational Linguistics*, Saarbrücken, Germany, (2000).

[12] J. Swales, *Genre Analysis. English in Academic and Research Settings*, Cambridge University Press, 1990.