Tochtermann, Maurer (Eds.): Proceedings of the I-KNOW '06, Graz
6th International Conference on Knowledge Management
Journal of Universal Computer Science, pp. 449-456, ISSN 0948-695x

Distinguishing Topic from Genre

Benno Stein and Sven Meyer zu Eissen

(Faculty of Media / Media Systems
Bauhaus University Weimar, Germany
benno.stein@medien.uni-weimar.de)

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?*

We present instruments to answer this question as well as to make a distinct statement about the homogeneity of a categorization.

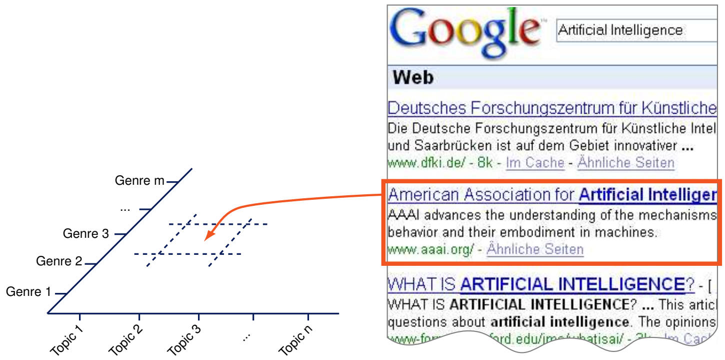
Key Words: Genre analysis, Personal IR, Knowledge discovery, Unsupervised learning
Category: H.3.3 Information Search and Retrieval, Retrieval models, Information filtering

1 Introduction and Background

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. The focus of this paper is on text documents, and, in this connection genre describes 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 [see 1], but, for the time being it is sufficient to remember the following characterization: Genre and topic are orthogonal—or, with Dewdney [Dewdney et al. 2001]: "The form is the substance."

Consequently, genre *identification* shall discover groups of texts that share a common form of transmission, purpose, or discourse properties [Santini 2004; Swales 1990]. If we imagine the "ideal retrieval process", an information seeker shall be able to circumscribe both the topic he or she is interested in and the preferred genre, such as research papers, personal experience reports, or commercial product information. And,

[1] Santini has compiled an up-to-date discussion of this term [Santini 2004].



Deutsches Forschungszentrum für Künstliche
Die Deutsche Forschungszentrum für Künstliche Intel
und Saarbrücken ist auf dem Gebiet innovativer ...
www.dfki.de/ - 8k - Im Cache - Ähnliche Seiten

American Association for **Artificial Intelliger**
AAAI advances the understanding of the mechanisms
behavior and their embodiment in machines.
www.aaai.org/ - Ähnliche Seiten

WHAT IS **ARTIFICIAL INTELLIGENCE**? - [
WHAT IS **ARTIFICIAL INTELLIGENCE**? **...** This articl
questions about **artificial intelligence**. The opinions

Figure 1: The ideal retrieval process? Text snippets from result lists are organized and presented in two dimensions, topic and genre.

instead of compiling a list with search results, documents may be organized within a topic dimension and a genre dimension (see Figure 1). Other retrieval scenarios that could benefit from an automatic genre identification are: The search in a company's intranet where documents fall into genre categories like "invoice", "dunning letter", or "customer presentation", while topics may cover product information, market surveys, and others. Likewise, educational material may be retrieved either with respect to its type (genre) such as "exercise", "reading', or "tutorial", but also with respect to the treated topic. Finally, note that meta-knowledge about the classification type can be exploited to construct a better classifier for a given document collection.

The remainder of this section gives an overview over related work and outlines the typical concepts of retrieval models for genre classification. The main contribution of this paper is Section 2, where we develop the necessary instruments to provide an answer to the question posed at the outset.

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:

1. definition of particular genre classes,
2. compilation of a respective genre corpus,
3. 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 [Bretan et al. 1999]. Based on an explorative 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 [Roussinov et al. 2001]. 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 [Dimitrova et al. 2002]. 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 [Lee and Myaeng 2002]. 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 [Meyer zu Eißen and Stein 2004]. 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 [Boese and Howe 2005].

1.2 Retrieval Models for Genre Classification

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 [Stamatatos et al. 2000; Lee and Myaeng 2002], classifiers that employ linguistic features along with additional features related to text statistics and computational linguistics [Kessler et al. 1997], or both [Finn and Kushmerick 2003; Meyer zu Eißen and Stein 2004]. 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 for genre identification can be built. By contrast, to capture the gist of a document with respect to its *topic*, the vector space model 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. In the following we use R_T and R_G to denote the underlying retrieval models for topic and genre respectively.

2 Identifying Categorization Types

Let D be a set of documents. An exclusive categorization $\mathcal{C} = \{C_1, \ldots, C_k\}, C_i \subseteq D$, is a division of D into sets for which the following conditions hold:

- $\bigcup_{C_i \in \mathcal{C}} C_i = D,$ 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 considerations, by genre considerations, or by both. Rendered precisely: For a categorization $\mathcal{C}$ let $\tau(C)$, $C \in \mathcal{C}$, designate the type (either topic or genre) of the document set C. Then, $\mathcal{C}$ is called a homogeneous or a "pure" (topic or genre) categorization, if $\forall C_i, C_j \in \mathcal{C} : \tau(C_i) = \tau(C_j)$. If $\tau(C_i) \neq \tau(C_j)$ for some $C_i, C_j \in \mathcal{C}$, a categorization is called inhomogeneous or "impure". Note that the type identification for a categorization $\mathcal{C}$ must take into account whether $\mathcal{C}$ is homogeneous or not. The next two subsections address the related situations, while Subsection 2.3 reports on interesting experiments.

2.1 Analyzing Homogeneous Categorizations

We identify a categorization's type by comparing the solutions of model fitting problems, $\Pi(R, \mathbf{D}, \theta)$, under different retrieval models R. $\mathbf{D}$ is the set of data points defined by $\mathcal{C}$, say, $\mathbf{D} = \{(d_i, j) \mid d_i \in (D \cap C_j), C_j \in \{C_1, \ldots, C_k\}\}$, and θ denotes a set of parameters to be fitted. The solution $\hat{\theta}$ of Π defines a predictor $\gamma_{R,\hat{\theta}} : D \to \{1, \ldots, k\}$.

Depending on the kind of fitting problem, $\hat{\theta}$ represents the parameters of a regression function, of a discriminant analysis, or of a neural network. The quality of the fit can be quantified with the F-measure; in particular, a perfect fit of the data points $\mathbf{D}$ means that the documents in D are classified without error and that the categorization $\mathcal{C}$ is exactly resembled. $\Pi(R, \mathbf{D}, \theta)$ is solved for both retrieval models, $R = R_T$ and $R = R_G$, and the categorization type "topic" is assumed for $\mathcal{C}$ if under R_T a higher F-measure value can be achieved than under R_G. Otherwise, the categorization type "genre" is assumed.

In our experiments a cluster analysis in the form of k-means is applied to solve the model fitting problems $\Pi(R_T, \mathbf{D}, \theta)$ and $\Pi(R_G, \mathbf{D}, \theta)$. I. e., a solution $\hat{\theta}$ of Π defines k centroid vectors each of which characterizes a category $C \in \mathcal{C}$. The rationale is that an unsupervised analysis approach is superior to an supervised approach here, because, firstly, it takes advantage of the entire set $\mathbf{D}$, and, secondly, it copes better with small or unequally distributed categories [see 2]. The entire analysis comprises the following steps:

1. Construct for each $d \in D$ two document models, one under the topic document retrieval model, R_T, and one under the genre document retrieval model, R_G.

2. Based on a similarity measure, Euclidean similarity or cos-similarity, construct two similarity graphs G_T and G_G. The edge weights in these graphs result from the similarity computations under R_T and R_G respectively.

3. Model fitting. Apply the k-means clustering algorithm to the graphs G_T and G_G. The resulting clusterings are designated as $\mathcal{C}_T$ and $\mathcal{C}_G$.

4. Compute the F-measure to quantify the quality of the fit between $\mathcal{C}$ and $\mathcal{C}_T$ as well as between $\mathcal{C}$ and $\mathcal{C}_G$ [see 3]. The resulting values are designated as $F_{\mathcal{C}_T}$ and $F_{\mathcal{C}_G}$.

[2] At heart we don't know anything about the nature of the distribution and prepare for the worst.

[3] Note that also another external reference measure can be applied.

5. Analyze $|F_{\mathcal{C}_T} - F_{\mathcal{C}_G}|$, the difference in the fitting quality. If $|F_{\mathcal{C}_T} - F_{\mathcal{C}_G}|$ is significant, $\mathcal{C}$ is organized under topic considerations if $F_{\mathcal{C}_T} > F_{\mathcal{C}_G}$, and under genre 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 $\mathcal{C}' = \{C'_1, \ldots, C'_l\}$. It combines the recall and precision statistics, where 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:

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

A perfect clustering matches the given categories exactly and yields an F-measure value of 1; an F-measure value close to zero indicates a high model fitting error and very little congruence.

2.2 Analyzing Impure Categorizations

The previously introduced analysis approach will come to its limits if a larger part of a categorization $\mathcal{C}$ is from a different type than the rest; Subsection 2.3 reports on an experiment to illustrate this effect.

We now introduce an instrument for both to decide whether two categories C_i, C_j in a categorization $\mathcal{C}$ are of the same type and to quantify the overall degree of $\mathcal{C}$'s homogeneity. Basic idea is the construction of a similarity graph for the document collection D, measuring its average similarity under some retrieval model R, and analyzing the increase or decline of the similarity for single categories $C \in \mathcal{C}$ under this retrieval model. The idea can be operationalized using a measure of *relative density*, ρ, which is introduced next.

A graph $G = \langle V, E, w \rangle$ is called sparse if $|E| = \mathcal{O}(|V|)$; it is called dense if $|E| = \mathcal{O}(|V|^2)$. Put another way, we can compute the density exponent α for a graph from the equation $|E| = |V|^\alpha$. With $w(G) := |V| + \sum_{e \in E} w(e)$, this relation extends naturally to weighted graphs [see 4]:

$$w(G) = |V|^\alpha \quad \Leftrightarrow \quad \alpha = \frac{\ln\big(w(G)\big)}{\ln\big(|V|\big)}$$

Obviously, α can be used to compare the density of an induced subgraph $G_i = \langle V_i, E_i, w_i \rangle$ of G to the density of G, leading to the relative density, $\rho(G_i)$:

$$\rho(G_i) = \frac{w(G_i)}{|V_i|^\alpha}$$

[4] $w(G)$ denotes the total edge weight of G plus the number of nodes, $|V|$, which serves as adjustment term for graphs with edge weights from the interval [0, 1].

G_i is sparse compared to G if $\rho(G_i) < 1$; G_i is dense compared to G if $\rho(G_i) > 1$. This connection can be used to quantify a retrieval model's ability to capture the intrinsic similarity relations of the categories in $\mathcal{C}$.

Let $\mathcal{C}$ be a—possibly impure—categorization of a document set D, and let G_T and G_G denote the similarity graphs under the retrieval models R_T and R_G. For a category $C_i \in \mathcal{C}$ its relative density under both retrieval models computes as follows:

$$\rho_T(C_i) = \frac{w(G_{T_i})}{|V_i|^{\alpha_T}}, \quad |V|^{\alpha_T} = w(G_T), \qquad \rho_G(C_i) = \frac{w(G_{G_i})}{|V_i|^{\alpha_G}}, \quad |V|^{\alpha_G} = w(G_G)$$

Note that $\rho(C_i)$ quantifies the *retrieval-model-specific* similarity increase in category C_i, and hence $|\rho_T(C_i) - \rho_G(C_i)|$ quantifies the superiority of one retrieval model over the other. If $|\rho_T(C_i) - \rho_G(C_i)|$ is significant, C_i is assumed to comprise documents of a single topic if $\rho_T(C_i) > \rho_G(C_i)$, and of a single genre class otherwise.

Based on ρ the overall degree of the homogeneity of $\mathcal{C} = \{C_1, \dots, C_k\}$, here designated as homogeneity coefficient $\bar{\tau}(\mathcal{C})$, can be defined as follows:

$$\bar{\tau}(\mathcal{C}) = \frac{\left| \sum_{i=1}^{k} \left(\rho_T(C_i) - \rho_G(C_i) \right) \right|}{\sum_{i=1}^{k} |\rho_T(C_i) - \rho_G(C_i)|}$$

A value of $\bar{\tau}(\mathcal{C})$ close to 1 indicates a homogeneous categorization; a value of $\bar{\tau}(\mathcal{C})$ close to zero gives evidence to a highly impure categorization.

2.3 Experiment Setups and Results

As retrieval model for topic, R_T, the standard vector space model with the $tf \cdot idf$-weighting scheme is employed. The genre retrieval model, R_G, for a document d is a vector $\mathbf{d}$ comprising high-level features for style, part-of-speech, and various closed-class word sets. Discriminant analysis was used to select the most informative features (about 40) along with appropriate weighting schemes [see 5]. The similarity measure φ_G of R_G is defined as $\varphi_G(\mathbf{d}_i, \mathbf{d}_j) = e^{-||\mathbf{d}_i - \mathbf{d}_j||}$ [see 6].

Our experiments rely on two corpora, RCV1 and a special genre corpus of [Meyer zu Eißen and Stein 2004], 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. The following experiments were set up:

- Type identification of homogeneous categorizations dependent on $|D|$, the size of the document set. Based on the corpora, 40 categorizations of different sizes and under both topic and genre considerations were compiled and analyzed. It turned out that for each of these categorizations its type could be unambiguously identified

[5] Our R_G is comparable to the model in [Bretan et al. 1999], but introduces also new concepts.
[6] The cos-similarity cannot be applied since the angle between the vectors is from $[0; 2\pi]$.

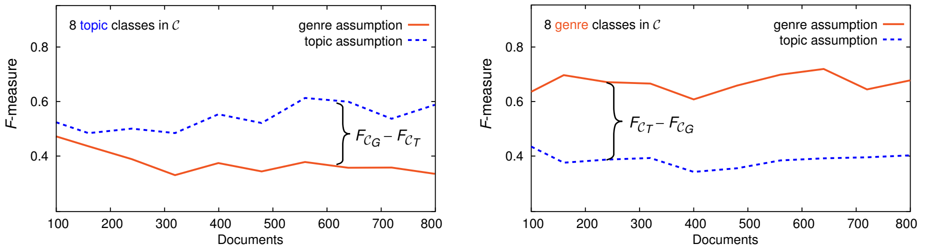


Figure 2: Type identification of homogeneous categorizations. The plots quantify the adequacy of the document models R_T and R_G. They unveil whether a categorization $\mathcal{C}$ is organized by topic (left) or by genre (right).

by computing $|F_{\mathcal{C}_T} - F_{\mathcal{C}_G}|$. Figure 2 shows the developing of the respective F-measure values $F_{\mathcal{C}_T}$ and $F_{\mathcal{C}_G}$. Given the similarity graphs, the runtime complexity is dominated by the cluster analysis, and, using k-means, linear in D.

– Type identification of homogeneous categorizations dependent on $|\mathcal{C}|$, the number of categories. As one may not have expected, the experiments show that the type identification are rather unaffected by the category number.

– Analysis of impure categorizations. The plot on the left-hand side in Figure 3 illustrates the connection between $|F_{\mathcal{C}_T} - F_{\mathcal{C}_G}|$ and the degree of impurity: As expected, the reliability of $|F_{\mathcal{C}_T} - F_{\mathcal{C}_G}|$ can only be guaranteed for categorizations $\mathcal{C}$ that are dominated by a single type, which meant more than 80% here. The plot on the right-hand side certifies the robustness of our technology: $\bar{\tau}$ turns out to be an ideal predictor of a categorization's degree of homogeneity.

3 Summary and Discussion

With *"Can topic be distinguished from genre?"* we have identified a new question, and we have introduced instruments to answer it. The experiments addressed both homogeneous and impure categorizations, and we could learn about the feasibility and robustness of our solutions. In this connection we used, among others, a measure of

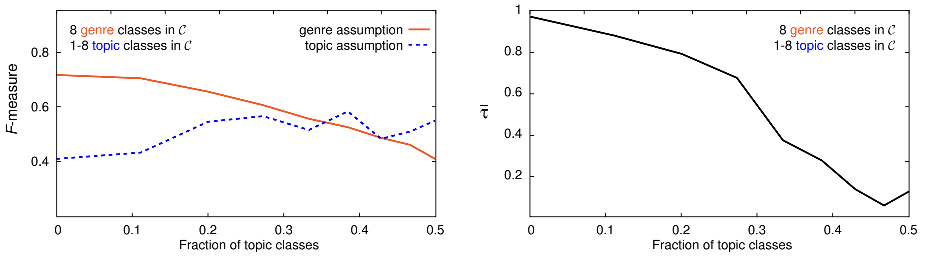


Figure 3: Analysis of impure categorizations. The left plot shows what happens to $|F_{\mathcal{C}_T} - F_{\mathcal{C}_G}|$ if a categorization becomes more and more impure. Note that the degree of impurity (or homogeneity) can be directly read off from the function $\bar{\tau}$, illustrated in the right plot.

relative density, which makes similarity graphs comparable with respect to their cohesiveness. This measure forms the basis of the homogeneity coefficient $\bar{\tau} : \boldsymbol{\mathcal{C}} \rightarrow [0;1]$, which predicts for a categorization $\mathcal{C} \in \boldsymbol{\mathcal{C}}$ its degree of homogeneity.

Note that only non-hierarchical categorizations were treated in this paper. However, if a complex hierarchy with several levels is given, type information can be analyzed and propagated bottom up, whereas all leafs of the same parent node are treated as one categorization $\mathcal{C}$.

Our current research continues the work of this paper and focuses on complexity issues: Compared to R_T the computation of R_G is much more expensive, and we are working on heuristic approaches to estimate reliable genre features.

References

Elisabeth Sugar Boese and Adele E. Howe. Effects of Web Document Evolution on Genre Classification. In *Proceedings of the CIKM'05*. ACM Press, November 2005.

Ivan Bretan, Johan Dewe, Anders Hallberg, and Niklas Wolkert. Web-specific genre visualization, 1999.

Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. The form is the substance: Classification of genres in text. In *Proceedings of ACL Workshop on HumanLanguage Technology and Knowledge Management*, 2001.

M. Dimitrova, A. Finn, N. Kushmerick, and B. Smyth. Web genre visualization. In *Proceedings of the Conference on Human Factors in Computing Systems*, 2002.

Aidan Finn and Nicholas Kushmerick. Learning to Classify Documents According to Genre. In *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. Automatic detection of text genre. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Somerset, New Jersey, 1997. Association for Computational Linguistics.

Yong-Bae Lee and Sung Hyon Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *Proc. 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150. ACM Press, 2002. ISBN 1-58113-561-0.

Sven Meyer zu Eißen and Benno Stein. Genre Classification of Web Pages: User Study and Feasibility Analysis. In Susanne Biundo, Thom Frühwirth, and Günther Palm, editors, *KI 2004: Advances in Artificial Intelligence*, volume 3228 LNAI of *Lecture Notes in Artificial Intelligence*, pages 256–269, Berlin Heidelberg New York, September 2004. Springer. ISBN 0302-9743.

Dmitri Roussinov, Kevin Crowston, Mike Nilan, Barbara Kwasnik, Jin Cai, and Xiaoyong Liu. Genre based navigation on the web. In *Proceedings of the 34th Hawaii International Conference on System Sciences*, 2001.

Marina Santini. State-of-the-Art on Automatic Genre Identification. Technical report, ITRI, University of Brighton, UK, 2004.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text genre detection using common word frequencies. In *Proceedings of the 18th Int. Conference on Computational Linguistics*, Saarbrücken, Germany, 2000.

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