Stylistic Document Retrieval for Turkish

Daniya Zamalieva, Fırat Kalaycılar, Aslı Kale, Selen Pehlivan, Fazlı Can
Department of Computer Engineering
Bilkent University
06800 Bilkent Ankara, Turkey
Email: {daniya,firatk,akale,pselen,canf}@cs.bilkent.edu.tr

***Abstract*—In information retrieval (IR) systems, there are a query and a collection of documents compared with this query and ranked according to a particular similarity measure. Since texts with the same content can be written by different authors, the writing styles of the documents change as well accordingly. This observation brings the idea of investigating text by means of style. In this paper, we analyze text documents in terms of stylistic features of the written text and measure effectiveness of these features in an IR system. Our main focus is on Turkish text documents. Although there are many studies about broadening IR systems with style based enhancement, there is no similar application for Turkish which performs retrieval depending purely on style.**

I. Introduction

Document retrieval systems try to match given queries with a collection of unstructured text documents. In general, they present a list of documents ordered from most relevant to less relevant. Distances between the query and each document are computed using several features. One common feature is the number of indexing terms shared by the query and a document. This shows that classical retrieval systems are mainly based on the similarity measurements such as frequencies of common textual units like words, phrases, indexing terms etc.[1]

In this study, we concentrate on the style of documents rather than their contents. Style comprises the structural and syntactic choices of an author that are independent from the subjects of writings. This indicates the uniqueness of author in terms of linguistic tendencies. Since there is no formal definition for document style, firstly the investigation of candidate features representing document style was conducted.

Based on this notion, we present a Turkish document retrieval approach relying on stylistic features. In this system, we model stylistic tendencies in terms of certain measurements and propose a numerical representation corresponding to document style. Input of the system is a textual unit, preferably a query document or a query paragraph. During retrieval, in contrast to the classical methods, this system computes distances between the query and the documents using their numerical style representations. Therefore the resulting list contains documents with similar stylistic tendencies.

One possible use of our system can be discovery of the probable author of an anonymous text like poem, short story, and newspaper columns. For example, there are a lot of literary works whose authors are unknown. We suggest that by comparison of these anonymous works with the writings of certain authors in terms of style, the probable author can be determined. In addition, this system can serve the users who seek for documents that have resemblance in style with the one they are interested in. By this way, people can discover other authors which create literary works they would probably be attracted to.

The document collection we used consists of newspaper columns written by different authors. We expect that for a new document (query) written by an author who has several articles in the collection, our system will primarily retrieve the existing documents of that writer.

The rest of the paper is organized as follows. In Section II, a brief discussion of related work is given. Section III contains a detailed description of the retrieval approach and stylistic features. Section IV presents performance evaluation using a collection of newspaper columns. Finally, Section V provides a summary and lists the contributions.

II. Related Work

Most of the effort on information retrieval systems covers enhancement strategies to increase retrieval performance. A part of the research is based on methods that enhance retrieval by including topic information. These approaches put topic relevance into consideration during retrieval. On the other hand, [2] states that deviations among documents are not only topical but also stylistic.

Analyzing subjective values of a given text in IR systems is a multi-disciplinary area including linguistics and computer science. Generally, each author has his own characteristic writing style that is independent of the topic. This inspiration expands approaches to improve IR systems with stylistic features. [2] describes a prototype which introduces stylistic items used for measurements of document style. Measurements with respect to these items are then merged using non-parametric multivariate method, such as decision tree learning approach. Study is tested on The Wall Street Journal articles and materials from Internet. According to [2], stylistic items are divided into two sets: lexical statistics including average word length, long word counts, number of pronouns, number of digits, and syntactic statistics such as average length of a sentence.

Similar work is presented by Argamon et. al [3]. The main aim of the study is to find a computationally efficient formulation of linguistic features to classify text styles. Then, machine learning techniques are applied to build a model to discriminate styles.

978-1-4244-5023-7/09/$25.00 ©2009 IEEE

ISCIS September 14-16, 2009
METU
Northern Cyprus Campus

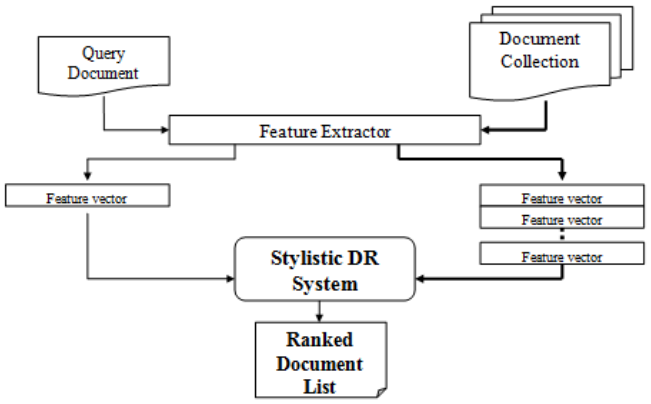

Fig. 1. System Overview.

Some other studies analyzing text style are based on author attribution [4], [5], [6], genre attribution [7] and semantic orientation [8]. For example, Argamon et al. [5] learn the models for discrimination of different authors by applying several multiclass variants of the Winnow [9] algorithm to the feature vectors corresponding to texts. Finn et al. [7] discusses and compares three different approaches for classification of documents by genre: bag of words techniques, part-of-speech statistics, and hand-crafted shallow linguistic features. Turney and Littman [8] introduce an algorithm for unsupervised learning of semantic orientation, the evaluative character of a word, from text. Similarly, Wiebe et al. [10] presents a corpus annotation project to investigate issues in manual annotation of private states in language such as opinions, emotions, sentiments, speculations, and evaluations.

Motivated by the previous research work discussed above and the lack of studies on document retrieval systems for Turkish, our work presents a style oriented method for Turkish IR systems.

III. Stylistic Document Retrieval

In this study, we want to represent each document in terms of a set of stylistic features. Firstly, the investigation of candidate features that can adequately describe the document style is performed. Details regarding the determined features, their properties, and feature extraction are presented further in this section.

The retrieval mechanism can be explained in the following way: first, the features are extracted from each document of the collection and stored. Then, the same features are extracted from a query document and its feature vector is compared with those obtained from the document collection. The similarity between two feature vectors can be calculated using the Mahalanobis distance [11]:

$$Similarity(\mathbf{x}_i, \mathbf{x}_j, \mathbf{H}) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \quad (1)$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are feature vectors, $\mathbf{H}$ is covariance matrix. For example, when $\mathbf{H}$ is chosen to be an identity matrix, the similarity measure corresponds to the Euclidean distance.

TABLE I
Measurements used for feature calculations.

Abbreviation	Description
$\mathcal{C}$	Number of conjunctions
$\mathcal{CS}$	Number of certainty suffixes
$\mathcal{CW}$	Number of certainty words
$\mathcal{FP}$	Number of formal pronouns
$\mathcal{FW}$	Number of formal words
$\mathcal{NA}$	Number of negative adjective/adverb
$\mathcal{P}$	Number of pronouns
$\mathcal{PA}$	Number of positive adjective/adverb
$\mathcal{Q}$	Number of questions
$\mathcal{S}$	Number of sentences
$\mathcal{SW}$	Number of stop words
$\mathcal{SYL}$	Number of syllables
$\mathcal{US}$	Number of uncertainty suffixes
$\mathcal{UW}$	Number of uncertainty words
$\mathcal{W}$	Number of words

However, since the Euclidean distance treats all feature components evenly, some problems can arise. The difference between feature components with high variance can dominate over those of other components and drastically affect the resultant similarity value. Therefore, in order to avoid this problem, each feature component should be normalized by using their variances. Hence, we define $\mathbf{H}$ as diagonal matrix whose entries are variances estimated from the document collection. When the Mahalanobis distance is employed, the feature component ranges are equalized.

After the similarity values between the query document and document collection are computed, the documents of the collection are ranked accordingly from the most similar to least similar. The overview of the implemented system is shown in Figure 1.

Below we discuss the features that are significant for style representation of documents. The measurements used for feature calculation and their abbreviations are listed in Table I.

Formality Measure: This feature defines a measure for the author's writing style in terms of formality. Our aim is to discover the structures that define formality of the document. It can be assumed that some authors tend to use formal expressions more frequently. Especially this is apparent when writer addresses the reader. In Turkish, some pronouns and some word suffixes can be used to identify formality measure of the texts.

The formality score is assigned according to the frequency of formal structures in the current document. In Turkish, *siz* pronoun can be used for politeness, addressing strangers, and showing respect. In addition, there are some words generated by adding a formal suffix *-iniz* that reflect formality. Based on these assumptions, the formality measure can be calculated as

TABLE II
SAMPLE ENTRIES OF TERM AND SUFFIX LISTS.

Positive Terms	Negative Terms	Certainty Terms	Uncertainty Terms	Conjunctions	Formality Suff.	Certainity Suff.	Uncertainty Suff.
güzel	çirkin	şüphesiz	belirsiz	ama	-iniz*	-malı*	-ebilir*
mutlu	mutsuz	bariz	sanki	fakat			
dostça	iğrenç	muhakkak	belki	lakin			
candan	berbat	mutlaka	muallak	ve			
cici	nankör	kesinlikle	meçhul	veya			
...	...	...	...	...			
431 terms	737 terms	12 terms	11 terms	50 terms	1 suffix	1 suffix	1 suffix

* and derivatives such as -ınız, -meli and -abilir.

$$Formality = \frac{1}{2}\left(\frac{\mathcal{FP}}{\mathcal{P}} + \frac{\mathcal{FW}}{\mathcal{W}-\mathcal{SW}}\right). \quad (2)$$

We formulated the equations by dividing the number of formal words in the document to the total number of words except the stop words. We discarded the number of stop words used in the document to reduce bias. This leads to a better evaluation of the text in terms of proposed features.

Positivity and Negativity Measure: We seek for a way to understand the general style of an author in terms of mood and emotion. We observed that some authors tend to favor positive words while others prefer to use negative words. Thus, we assume that evaluation of positivity and negativity in the whole text provides the idea about the general emotional tendency of the author.

Based on this assumption, the *positive (negative) terms list* that contains adjectives representing the positive (negative) attitude is constructed. The words are selected manually from the adjectives and adverbs published in the TDK web dictionary [12]. To provide objectivity, we select the words that are well-known as positive (negative) in Turkish. Refer to Table II for sample terms. Then, these lists are used to assign a positivity (negativity) score to the document as follows:

$$Positivity = \frac{\mathcal{PA}}{\mathcal{W}-\mathcal{SW}} \quad (3)$$

$$Negativity = \frac{\mathcal{NA}}{\mathcal{W}-\mathcal{SW}}. \quad (4)$$

Certainty and Uncertainty Measure: It can be observed that some authors prefer to write in a more certain (uncertain) manner than others. This attitude is expressed both by words and suffixes reflecting certainty (uncertainty). To assign a certainty (uncertainty) score based on the certain (uncertain) words usage, the *certainty (uncertainty) terms list* containing the words that represent the sureness and definiteness (unsureness and doubtfulness) is prepared and the frequency of occurrence of these terms in the current document is calculated as follows:

$$Certainty = \frac{1}{2}\left(\frac{\mathcal{CW}}{\mathcal{W}-\mathcal{SW}} + \frac{\mathcal{CS}}{\mathcal{W}-\mathcal{SW}}\right) \quad (5)$$

$$Uncertainty = \frac{1}{2}\left(\frac{\mathcal{UW}}{\mathcal{W}-\mathcal{SW}} + \frac{\mathcal{US}}{\mathcal{W}-\mathcal{SW}}\right). \quad (6)$$

Question Usage Measure: Some authors prefer to develop their writings by interacting with readers by asking questions. Since questions can be located by question marks, they are counted in the text to calculate the number of questions:

$$QuestionUsage = \frac{\mathcal{Q}}{\mathcal{S}} \quad (7)$$

Stop Word Usage Measure: The general tendency in text analysis is to eliminate stop words prior to processing the documents. However, usage of stop words in the text can give a clue about author style. For example, some authors use many stop words in their texts while others avoid using them frequently. We obtain the *Turkish stop word list* containing 114 words from [13]. By using the list, the occurrence frequency of these terms in the document is calculated by:

$$StopWordUsage = \frac{\mathcal{SW}}{\mathcal{W}} \quad (8)$$

Conjunction Usage Measure: Conjunctions are basic textual units that generally connect two phrases. Hence, the overuse of conjunctions implies complex and long sentences that are also distinctive features of document style. By using the *conjunction term list* from TDK web dictionary [12], the measure is evaluated as

$$ConjunctionUsage = \frac{\mathcal{C}}{\mathcal{W}} \quad (9)$$

Word Length Measure: Since word length measure is widely used in document style analysis, we include it in our feature set. We measure the average length of the words in terms of syllable as follows:

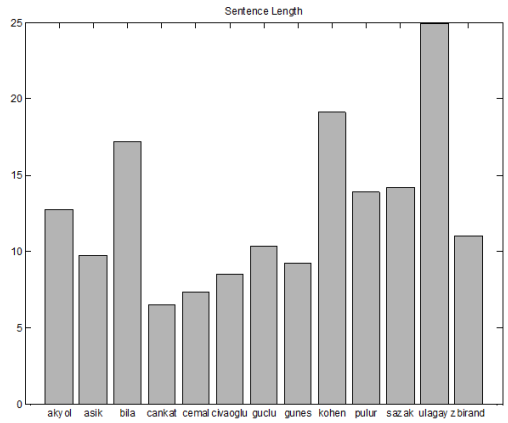

Fig. 2. Average Sentence Length of Documents Written by Different Authors

$$WordLength = \frac{\mathcal{SYL}}{\mathcal{W}}. \tag{10}$$

Sentence Length Measure: A significant property revealing stylistic characteristics of a document is the average length of the sentences. We suggest that this measure is adequate for style discrimination. For example, long sentences generally correspond to complex expressions whereas shorter ones convey simpler ideas. The formulation of this measure is given as

$$SentenceLength = \frac{\mathcal{W}}{\mathcal{S}}. \tag{11}$$

IV. Experimental Work

A. Document Collection

We evaluate the proposed system by using the document collection that consists of Turkish newspaper articles written by different authors. There are 13 authors and 60 articles for each. The authors are Taha Akyol, Melih Aşık, Fikret Bila, Berrin Cankat, Hasan Cemal, Güneri Civaoğlu, Abbas Güçlü, Hurşit Güneş, Sami Kohen, Hasan Pulur, Derya Sazak, Osman Ulagay, and Mehmet Ali Birand.

B. Author Style Analysis

When all authors are compared according to a stylistic feature component, we can examine that some stylistic features are discriminative for a subset of authors. For example, Osman Ulagay prefers to use longer sentences than Berrin Cankat as can be seen from Figure 2. Moreover, authors have different uncertainty degrees in their writings as illustrated in Figure 3. However, some feature components have similar values for most of the authors so they are not distinctive enough like word length shown in Figure 4.

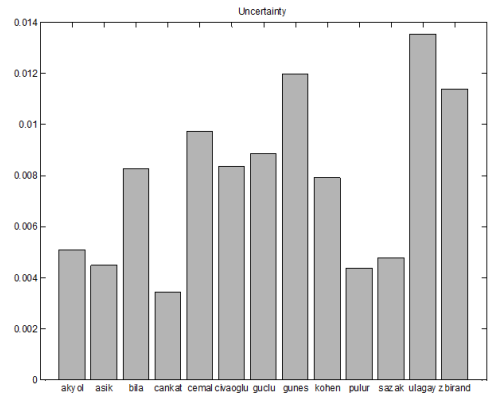

Fig. 3. Author vs. Average Value of Uncertainty

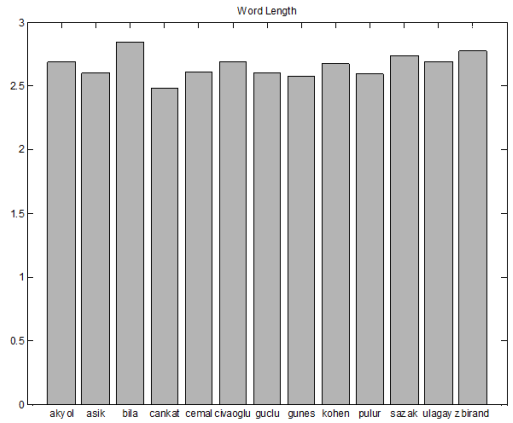

Fig. 4. Average Word Length of Documents Written by Different Authors

C. System Evaluation

The performance of the system is evaluated by precision measure. Precision is the percent of retrieved documents that are relevant to the query. For using this measure, the relevance information between a query document and the document collection is required. There is no available data set for Turkish stylistic document retrieval systems with appropriate groundtruth data. Thus, we assume that most authors tend to have their own style and two documents are considered as relevant if they are written by the same author.

We tested the system by using 780 query documents from all authors. After ranking the documents as explained in Section III, the corresponding precision values are calculated for the first 60 retrieved documents. By taking average of precision values of all these queries, we obtain the precision graph shown in Figure 5 (corresponding to solid black curve labeled as *all*). When we consider the first 5 and 10 retrieved documents, the precision of the system is about 0.84 and 0.77 respectively. These results show that although document style is a fuzzy and subjective concept, the performance achieved is high. Thus, it can be concluded that chosen features are successful enough to capture document style. Notice that the lowest precision obtained is 0.53.

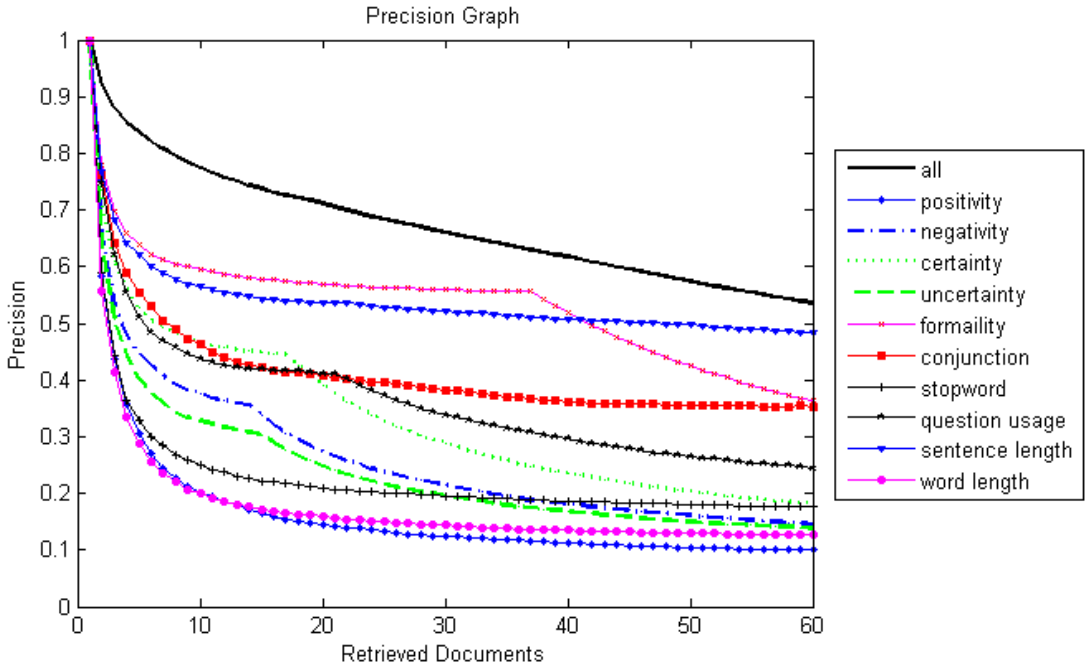

Fig. 5. Number of Retrieved Documents vs. Average Precision

In addition, we perform retrieval based solely on each feature component to investigate the effectiveness of it. The precision corresponding to each feature is presented on Figure 5. It is obvious that none of the feature components could outperform the combination of them. Besides this, we can infer that the most discriminative features are formality measure and sentence length. This confirms that authors tend to preserve the formality level and sentence length in their writings. In contrast, since word length and positivity features show low retrieval performance, we can conclude that they are less discriminative.

V. Conclusions

In this study, we develop a stylistic document retrieval approach for Turkish. We investigate the features that are able to capture the style of the document and use them for retrieval of the documents that are similar in style. Although generally style is used to enhance the existing retrieval systems, we implement a retrieval approach that is based purely on stylistic features. It is specialized for Turkish by construction of the term lists consisting of Turkish words.

As future work, we will examine other features which could enhance our retrieval performance. For example, usage of Ottoman Turkish words or slang existence can be investigated as new stylistic features. Moreover, the proposed system can be embedded to other retrieval systems in order to enhance their retrieval precision.

References

[1] G. Salton, "Another look at automatic text-retrieval systems," *Communications of the ACM*, vol. 29, no. 7, pp. 648–656, 1986.
[2] J. Karlgren, "Stylistic experiments for information retrieval," 2000.
[3] S. Argamon, C. Whitelaw, P. Chase, S. Hota, N. Garg, and S. Levitan, "Stylistic text classification using functional lexical features," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 6, pp. 802–822, 2007.
[4] A. McEnery and M. Oakes, "Authorship studies/textual statistics." 2000.
[5] S. Argamon, M. Šarić, and S. Stein, "Style mining of electronic messages for multiple authorship discrimination: first results," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM New York, NY, USA, 2003, pp. 475–480.
[6] D. Madigan, A. Genkin, D. Lewis, S. Argamon, D. Fradkin, and L. Ye, "Author identification on the large scale," in *Proc. of the Meeting of the Classification Society of North America*, 2005.
[7] A. Finn, N. Kushmerick, and B. Smyth, "Genre classification and domain transfer for information filtering," *Lecture notes in computer science*, pp. 353–362, 2002.
[8] P. Turney and M. Littman, "Unsupervised learning of semantic orientation from a hundred-billion-word corpus," *Arxiv preprint cs.LG/0212012*, 2002.
[9] N. Littlestone, "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm," *Machine learning*, vol. 2, no. 4, pp. 285–318, 1988.
[10] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2, pp. 165–210, 2005.
[11] P. Mahalanobis, "On the generalized distance in statistics," in *Proceedings of the National Institute of Science of India*, vol. 12, no. 1, 1936, pp. 49–55.
[12] TDK. Türk dil kurumu. [Online]. Available: http://www.tdk.gov.tr
[13] Turkish stopwords. [Online]. Available: http://www.ranks.nl/stopwords/turkish.html