Evaluation of Web Page Representations by Content Through Clustering*

Arantza Casillas[1], Víctor Fresno[2],
M. Teresa González de Lena[2], and Raquel Martínez[2]

[1] Dpt. Electricidad y Electrónica. UPV-EHU
arantza@we.lc.ehu.es
[2] Dpt. Informática, Estadística y Telemática, URJC
{v.fresno,m.t.gonzalez,r.martinez}@escet.urjc.es

Abstract. In order to obtain accurate information from Internet web pages, a suitable representation of this type of document is required. In this paper, we present the results of evaluating 7 types of web page representations by means of a clustering process.

1 Web Document Representation

This work is focused on web page representation by text content. We evaluate 5 representations based solely on the plain text of the web page, and 2 more which in addition to plain text use HTML tags for emphasis and the "title" tag. We represent web documents using the vector space model. First, we create 5 representations of web documents which use only the text plain of the HTML documents. These functions are: Binary (B), Term Frequency (TF), Binary Inverse Document Frequency (B-IDF), TF-IDF, and weighted IDF (WIDF). In addition we use 2 more which combine several criteria: word frequency in the text, the words appearance in the title, positions throughout the text, and whether or not the word appears in emphasized tags. These representations are the Analitic Combination of Criteria (ACC) and the Fuzzy Combination of Criteria (FCC). The first one [Fresno & Ribeiro 04] uses a linear combination of criteria, whereas the second one [Ribeiro et al. 03] combines them by using a fuzzy system.

2 Experiments and Conclusions

We use 3 subsets of the BankSearch Dataset [Sinka & Corne] as the web page collections to evaluate the representations: (1) ABC&GH is made up of 5 categories belonging to 2 more general themes; (2) G&H groups 2 categories that belong to a more general theme; and (3) A&D comprises 2 separated categories. Thus, the difficulty of clustering the collections is not the same. We use 2 feature reduction methods: (1) considering only the terms that occur more than a minimum times ("Mn", 5 times); (2) removing all features that appear in more than x documents ("Mx", 1000 times). For ACC and FCC we use the proper

* Work supported by the Madrid Research Agency, project 07T/0030/2003 1.

A. Apostolico and M. Melucci (Eds.): SPIRE 2004, LNCS 3246, pp. 129–130, 2004.

Table 1. Clustering results with the different collections and representations

	ABC&GH				G&H				A&D			
Represent.	**N. Feat.**	**F-me.**	**Entr.**	**T. s.**	**N. Feat.**	**F-me.**	**Entr.**	**T. s.**	**N. Feat.**	**F-me.**	**Entr.**	**T. s.**
ACC (10)	5,188	0.805	0.175	26	3,802	0.891	0.149	7	2,337	0.988	0.026	4
ACC (7)	4,013	0.803	0.176	18	2,951	0.869	0.168	5	1,800	0.988	0.026	3
ACC (5)	3,202	0.763	0.184	16	2,336	0.888	0.152	4	1,409	**0.989**	**0.025**	3
ACC (4)	2,768	0.818	0.170	13	1,999	**0.898**	**0.143**	4	1,228	**0.989**	**0.025**	2
FCC (10)	5,620	0.959	0.071	34	3,933	0.879	0.153	8	2,580	0.974	0.048	4
FCC (7)	4,114	0.952	0.080	19	2,813	0.851	0.167	5	1,886	0.972	0.051	3
FCC (5)	3,076	0.951	0.082	15	2,047	0.831	0.176	4	1,422	0.978	0.044	2
FCC (4)	2,544	0.955	0.077	11	1,654	0.823	0.194	3	1,188	0.972	0.051	2
B(Mn-Mx)	12,652	0.960	0.073	85	11,175	0.667	0.272	24	4,684	0.985	0.089	9
B(Mn)	13,250	0.963	0.066	61	11,499	0.774	0.228	31	4,855	0.975	0.045	8
B-IDF(Mn-Mx)	12,652	0.976	0.047	80	11,175	0.740	0.247	22	4,684	0.982	0.039	9
B-IDF(Mn)	13,250	**0.979**	**0.043**	65	11,499	0.814	0.202	30	4,855	0.974	0.048	9
TF(Mn-Mx)	12,652	0.938	0.096	89	11,175	0.775	0.230	23	4,684	0.975	0.046	8
TF(Mn)	13,250	0.937	0.095	62	11,499	0.856	0.178	30	4,855	0.953	0.073	8
TF-IDF(Mn-Mx)	12,652	0.466	0.255	91	11,175	0.858	0.176	21	4,684	0.982	0.034	9
TF-IDF(Mn)	13,250	0.966	0.062	62	11,499	0.880	0.159	28	4,855	0.975	0.037	11
WIDF(Mn-Mx)	12,652	0.907	0.127	88	11,175	0.771	0.230	22	4,684	0.905	0.136	9
WIDF(Mn)	13,250	0.924	0.111	69	11,499	0.776	0.228	29	4,855	0.916	0.114	9

weighting function of each one as the reduction function, by selecting the n most relevant features on each web page (i. e. ACC(4) means that only the 4 most relevant features of each page are selected). Notice that only B, TF, ACC and FCC are independent of the collection information. A good representation is one which leads to a good clustering solution. Since we work with a known, small number of classes (2 in these collections) we use a partition clustering algorithm of the CLUTO library [Karypis]. We carry out an external evaluation by means of F-measure and entropy measures.

The results can be seen in Table 1. It shows the number of features, the values of the external evaluation and the time taken in the clustering process. The experiments show that no single representation is the best in all cases. ACC is involved in the best results of 2 collections and the results of FCC are similar or, in some cases, better than with the others. These results suggest that using light information from the HTML mark-up combined with textual information leads to good results in clustering web pages. The ACC representation optimizes the web page's representation using less terms, and does not need collection information.

References

[Fresno & Ribeiro 04] Fresno, V., Ribeiro, A.: "An Analytical Approach to Concept Extraction in HTML Environments". *IIIS*. Kluwer A. Pub., (2004) 215-235.

[Karypis] Karypis G. "CLUTO: A Clustering Toolkit". Technical Report: 02-017. University of Minnesota, Department of Computer Science, Minneapolis, MN 55455.

[Ribeiro et al. 03] Ribeiro, A., Fresno, V., García-Alegre, M., and Guinea, D.: "A Fuzzy System for the Web page Representation". *Intelligent Exploration of the Web*, Springer-Verlag, (2003) 19-38.

[Sinka & Corne] Sinka, M. P., Corne, D. W. BankSearch Dataset. `http://www.pedal.rdg.ac.uk/banksearchdataset/`