As We May Perceive: Finding the Boundaries of Compound Documents on the Web

Pavel Dmitriev
Cornell University, Ithaca, NY 14853
dmitriev@cs.cornell.edu

ABSTRACT

This paper considers the problem of identifying on the Web *compound documents* (*cDocs*) – groups of web pages that in aggregate constitute semantically coherent information entities. Examples of cDocs are a news article consisting of several html pages, or a set of pages describing specifications, price, and reviews of a digital camera. Being able to identify cDocs would be useful in many applications including web and intranet search, user navigation, automated collection generation, and information extraction.

In the past, several heuristic approaches have been proposed to identify cDocs [1][5]. However, heuristics fail to capture the variety of types, styles and goals of information on the web, and do not account for the fact that the definition of a cDoc often depends on the context. This paper presents an experimental evaluation of three machine learning-based algorithms for cDoc discovery. These algorithms are responsive to the varying structure of cDocs and adaptive to their application-specific nature. Based on our previous work [4], this paper proposes a different scenario for discovering cDocs, and compares in this new setting the local machine learned clustering algorithm from [4] to a global purely graph based approach [3] and a Conditional Markov Network approach previously applied to noun coreference task [6]. The results show that the approach of [4] outperforms the other algorithms, suggesting that global relational characteristics of web sites are too noisy for cDoc identification purposes.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, retrieval models.* I.2.6 [**Artificial Intelligence**]: Learning.

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

WWW, Compound Documents, Machine Learning.

1. INTRODUCTION

This paper discusses the problem of automatically identifying the boundaries of information resources on the Web, focusing on a specific type of resources called *compound documents (cDocs)* [1][4]. A compound document is a set web pages that in aggregate correspond to a semantically coherent information entity. A typical example of a cDoc is a web news article consisting of several html pages[1], or a set of web pages describing a digital camera, with different pages dedicated to specifications, reviews, photographic tests, etc[2].

The ability to automatically identify cDocs would be useful in many applications, such as web and intranet search, improved usability and functionality of web applications, automated collection generation, information extraction and content summarization. For example, applied to web search it would let search engines index a document as a whole, instead of its individual pages, which could lead to improvement in the quality of search results [1], more accurate link analysis algorithms, and more accurate presentation of the results.

The problem of identifying cDocs was previously addressed in [1] and [5] using heuristic approaches. However, these approaches fail to take into account the variety of goals with which web sites are created leading to different conventions being used in different user domains. These approaches also do not account for the application-specific nature of cDocs[3], enforcing a one-size-fits-all approach.

Our previous work [4] proposed a machine learning based approach to identifying cDocs, designed to address the above problems. The algorithm used as training data several web sites with manually labeled cDocs, and was then applied to identify cDocs on new web sites. Experiments on a set of educational web sites showed promising results, but also revealed several problems with this approach (see [4] for details).

This paper proposes a different scenario for identifying cDocs. In this scenario, several example cDocs from a web site are used to train the algorithms that will then identify all other cDocs on the same web site. This paper compares for this new scenario the algorithm of [4] to two other algorithms previously used in graph mining and natural language processing domains.

2. ALGORITHMS

Baseline. Several approaches based on the heuristics from [1] and [5] were used as a baseline. One heuristic rule found to be very useful in [1] was to split the web site into cDocs according to the directory structure. Another approach is to use a single feature, e.g. content similarity, and compute its value for every pair of connected pages on the web site. Then, any threshold value t defines a binary relation R_t on the set of all web pages, where for two pages p_1 and p_2, they are in the relation $R_t(p_1,p_2)$ iff content similarity between p_1 and p_2 is greater than t. Transitive closure of this relation gives a set of cDocs. We tried 5 such approaches based on different features, using the optimal value of t in the experiments.

Weighted Graph Clustering. The *Weighted Graph Clustering*, or *WGC* [4], learns clustering of the web site based on a detailed analysis of the features of individual web pages and their immediate neighbors. On the training phase, the user provides a few examples of cDocs on a web site. Then, for every hyperlink at least one end of which is in a user-provided cDoc, a vector of feature values X_{ij} is computed. Given training data in the form of pairs (X_{ij}, *within*/*between*), where the label indicates whether the hyperlink is within a user-provided cDoc, a logistic regression model is trained to estimates the probability of the hyperlink being within a cDoc. On the inference phase, the learned model is applied to every hyperlink to generate for the web site a weighted graph. Then, a variant of the shortest link

[1] http://www.nytimes.com/2004/10/12/politics/campaign/12policy.html
[2] http://www.dpreview.com/reviews/canoneos40d/

[3] See [2] for a further discussion of this point.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

clustering algorithm, with automatically determined stopping criteria, is applied. The resulting clusters are taken as cDocs.

Collective Clustering. The *Collective Clustering*, or *CCL*, is a Conditional Markov Network model similar to the model of identity uncertainty proposed in [6] for the noun coreference task. Let P be a set of random variables corresponding to the pages of a web site, and let L be a set of binary random variables corresponding to hyperlinks and specifying whether the hyperlink is within or between cDocs. The CCL models the conditional probability distribution $P(L \mid P)$ using potential functions f_n over cliques in the web site graph:

$$P(L \mid P) = \frac{1}{Z_P} \exp\left(\sum_{i,j,n} \lambda_n f_n(p_i, p_j, l_{ij})\right)$$

The model is trained using the procedure described in [6]. The resulting labelling is taken as cDocs[4]. Note that even though only potential functions over pairs of pages (cliques of size 2) are used, the labelling decision for a certain pair influences the values of potential functions of the overlapping pairs. This collective nature of label assignment is the main difference of the CCL from the WGC, which makes labelling decisions independently.

Generalized Pattern Matching. The *Generalized Pattern Matching*, or *GPM* [3], does not analyze content of web pages, but performs a global analysis of the web site graph using pattern matching techniques. Based on the subgraphs corresponding to the user provided examples of cDocs the algorithm identifies their structural signatures (*cores*). Intuitively, cores are subgraphs often encountered inside cDocs which do not cross cDoc boundaries. Then, the GPM finds all occurrences of cores on the web site, and expands them to obtain complete cDocs according to a specified expansion rule. (see [2][3] for details).

It turned out that the original algorithm from [3] applied to the problem of finding cDocs suffers from severe performance problems. Therefore we used approximations on several steps of the algorithm, all based on sampling over search paths in the graph. Details can be found in [2].

3. EXPERIMENTS

The dataset consisted of 60 web sites from 3 categories collected mostly from DMOZ directory: 20 educational, 20 news, and 20 commercial web sites. On average, a web site had 169 pages and 1398 links. For every web site, cDocs were manually identified by 3 labelers, resulting in 19, 12, and 20 cDocs per web site on average. The average mutual agreements among labelers were rather low, between 0.4 and 0.6. This confirms the point mentioned earlier that the aggregation criteria vary depending on the user or application. It also suggests that one-size-fits-all heuristic approaches to cDoc identification will not be able to produce good results for all labelers.

For each web site, separate experiments were run for each labeler using 1, 2, and 3 cDocs as training examples, and all cDocs on the web site as a test set. Recall was used as a primary evaluation measure[5]. Tables 1 and 2 summarize the baseline performance. As one can see, the relative order of the methods is similar for all three labelers. There is also no single feature that does best for all three categories of web sites. The results suggest that no single feature is sufficient for identifying cDocs.

Table 1. Performance of baseline approaches on all sites.

	content	outlinks	filename	title	directory
L_1	0.23	0.28	**0.31**	**0.31**	0.26
L_2	0.31	**0.41**	0.38	0.34	0.13
L_3	0.24	0.28	**0.32**	0.28	0.11

Table 2. Performance of baseline approaches on different categories of sites for labeler L_1.

	content	outlinks	filename	title	directory
edu	0.2	**0.34**	0.22	0.32	0.25
news	0.24	0.18	**0.47**	0.3	0.07
com	0.25	0.32	0.25	0.3	**0.37**

Table 3. Performance of all algorithms on all sites.

	Best baseline	GPM	CCL	WGC
L_1	0.31	0.35(0.07)	0.34(0.1)	**0.69(0.06)**
L_2	0.41	0.39(0.08)	0.42(0.1)	**0.68(0.07)**
L_3	0.32	0.3(0.07)	0.33(0.06)	**0.66(0.06)**

Table 4. Performance of all algorithms on different categories of sites for labeler L_1 using 3 training cDocs.

	Best baseline	GPM	CCL	WGC
edu	0.34	0.41(0.03)	0.26(0.11)	**0.47(0.1)**
news	0.47	0.31(0.04)	0.39(0.11)	**0.85(0.04)**
com	0.37	0.34(0.09)	0.46(0.1)	**0.76(0.05)**

Tables 3 and 4 present the results for all algorithms using 3 training cDocs. The results are averages over 10 runs of the experiment, the numbers in parentheses are standard deviations. For all labelers and all categoris the WGC outperforms all other approaches. This indicates that, while no single feature can reliably identify cDocs, a machine learned combination of features can do that quite accurately. Poor performance of the GPM is due to a very high density of news and commercial web sites (on more sparse educational web sites the GPM performs significantly better). The reasons for poor performance of the CCL are not clear. One hypothesis is that the presence of meaningless navigational links among cDocs leads to imprecise parameter estimation, another hypothesis is that the approximation algorithm used for parameter estimation of the model is not appropriate for our problem. Developing a higher quality relational model for identifying cDocs is our primary direction for future work.

4. ACKNOWLEDGEMENTS

The author would like to thank Carl Lagoze, Thorsten Joachims, William Y. Arms, and Paul Ginsparg for valuable comments on this work, and Ryan Workman and Stuart Tettemer for help with labeling the data. This work was supported by the National Science Foundation grant number IIS-0430906.

5. REFERENCES

[1] Eiron, N., McCurley, K. S. Untangling compound documents on the web. In Proceedings of Hypertext'2003.

[2] Dmitriev, P. As We May Perceive: Finding the Boundaries of Compound Documents on the Web. Ph.D. Dissertation, Cornell University, January 2008.

[3] Dmitriev, P., Lagoze, C. Mining Generalized Graph Patterns based on User Examples. In Proceedings of ICDM'2006.

[4] Dmitriev, P., Lagoze, C., Suchkov, B. As We May Perceive: Inferring Logical Documents from Hypertext. In Proceedings of Hypertext'2005.

[5] Li, W.-S., Kolak, O., Vu, Q., Takano, H. Defining logical domains in a Web Site. In Proceedints of Hypertext'2000.

[6] McCallum, A., Wellner, B. Toward Conditional Models of identity uncertainty with application to proper noun coreference. In Proceedings of IJCAI-IIWeb'2003.

[4] The same approach as in [6] was used to prevent inconsistent labelings. Thus, a resulting labeling is always a non-overlapping set of clusters.

[5] Since the web sites were only crawled up to a certain depth, they could contain incomplete cDocs, making it difficult to use precision as an evaluation measure. Note, however, that since cDocs do not overlap it is not possible to optimize recall at the expense of precision in our case.