Clustering Web Documents with Tables for Information Extraction

Kostyantyn Shchekotykhin, Dietmar Jannach and Gerhard Friedrich
University Klagenfurt, Universitätsstraße 65-67, Klagenfurt, Austria
{kostya,dietmar,gerhard}@ifit.uni-klu.ac.at

ABSTRACT

One of the common approaches to extracting high-quality knowledge from Web sources is to exploit the redundancy of the published information. Therefore, a Web Mining System not only has to search for relevant Web pages but also has to somehow determine whether two pages describe the same entity in order to extract as much knowledge as possible about it. It has been shown that statistical clustering techniques are in general a suitable means to achieve this task by grouping documents that are supposed to contain similar information. However, when data is given in tabular form - which is for instance a typical way of describing items in online shops - existing document clustering algorithms show limited performance as documents containing tabular descriptions typically share a very common set of tokens although they describe different entities. In this paper we therefore propose a new document clustering approach that exploits hyperlinks and document metadata to extract candidates for entity names. These candidate names are subsequently used to cluster the documents and further improve these names, which are finally used to determine whether two documents describe the same entity. The detailed evaluation of our approach in two popular example domains showed its high accuracy in terms of precision and recall (F-Measure > 0.9).

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Clustering

General Terms

Algorithms, Experimentation

1. PROBLEM DESCRIPTION

Statistics-based Web Mining Systems (WMS) that exploit the redundancy of data on the Web for Information Extraction purposes have to first identify sets of Web documents that describe the same piece of information. This task of grouping the retrieved documents is commonly referred to as clustering. Existing approaches to document clustering are designed to work with general sources written in natural language, i.e., they analyze documents or their parts "as is". If, however, the desired information is described in tabular form, which is a common form of representing data on the Web in compact form, the applicability of these clustering methods is limited: Tables that describe *different* instances typically contain very common sets of words (like attribute names) and almost no grammar. On the other hand, it often happes that there are two documents that describe the same entity but use non-overlapping sets of tokens, which means that standard clustering techniques would put them into different groups.

The process of grouping tabular descriptions that describe the same entity can thus not be accomplished by the direct application of these existing methods. Therefore, it would be better to cluster documents containing tabular descriptions indirectly by means of some previously extracted unambiguous entity identifiers, like an *entity name*. These extracted names could consequently also be used to retrieve additional documents from the Web, as to increase redundancy and to improve the results of a statistics-based Web mining process. Unfortunately, such names or other identifiers are usually not explicitly given and there is no general rule where such identifiers typically appear on a Web page.

2. PROPOSED METHOD

Figure 1 summarizes a new clustering algorithm which is implemented in the ALLRIGHT Web Mining System. The algorithm accepts as an input a set of *candidates*, i.e., a set of *annotated* Web documents retrieved by the crawler component of the ALLRIGHT system which are supposed to contain the desired instance information. As an output, a set of clusters is returned; the calculation of the clusters proceeds in four main phases as follows.

In the PREPARE step, unnecessary information like punctuation or stop words are removed from the documents. Next, (*for-loop*), the X-MEANS [1] clustering technique is applied on each candidate in order to determine the most promising *identifier* for it. From the resulting clusters, we SELECT

K-CAP'07, October 28–31, 2007, Whistler, British Columbia, Canada.
ACM 978-1-59593-643-1/07/0010.

```
CLUSTERING ( in:candidates out:clusters)
{
  PREPARE( candidates )
  for each (candidate ∈ candidates)
  {
    nameClusters := X-MEANS( candidate )
    cluster := SELECT( nameClusters )
    candidate.identifier := TRANSFORM( cluster )
  }
  clusters := PARTITION( candidates )
  while (existIncoherentClusters( clusters ) )
  {
    cluster := removeMostIncoherentCluster( clusters )
    REMOVECORECOMPONENT(cluster)
    improvedClusters := PARTITION( cluster )
    clusters := clusters ∪ improvedClusters )
  }
  VALIDATE-AND-IMPROVE( clusters )
  return candidateClusters
}
```

Figure 1: ALLRIGHT Clustering algorithm

the most promising cluster by analyzing unused data like the contents of other tags that were not examined by X-Means. The selection method first calculates weights for each of the clusters. The weights are calculated as a ratio of the number of those documents, in which *all* tokens of an examined cluster can be found , to the number of all documents in the whole data set. Then, the algorithm calculates a centroid for all cluster scores and returns a cluster nearest to the centroid in the terms of Euclidian distance. Within the TRANSFORM method, the tokens of the chosen identifier are put back to the correct order.

After each candidate is associated with an identifier, a partitioning algorithm (PARTITION) is applied to produce clusters of candidates. The algorithm uses the *string metric* [2], which is defined on the interval $[0, 1]$, to measure the similarity of candidate identifiers.

The created clusters are then checked for coherency ("while-loop" in the algorithm). With coherency we mean that all elements in a cluster are similar to each other. If a cluster is incoherent, a core candidate is removed (REMOVE-CORE-COMPONENT) and the remaining candidates are analyzed by the partitioning algorithm once again.

Finally, a group improvement algorithm (VALIDATE-AND-IMPROVE) is subsequently used to identify groups that can be merged or split. The refinement step is based on checking - through additional search engine queries and a special *context similarity coefficient* - how often the names of groups can be found in the same document.

3. EVALUATION AND DISCUSSION

For evaluation purposes, we have tested the ALLRIGHT system in different domains. In one test run, the system's crawling component has in a first step automatically located and downloaded 3135 observations of digital cameras from 18 Web sites and 2930 observations from 12 Web sites for the domain of notebooks. These sets of documents were then analyzed with our new name recognition and clustering technique which determines those groups of documents that describe the same camera or notebook respectively.

Domain	Precision	Recall	F-Measure
Digital Camera	0.964	0.965	0.964
Notebook	0.943	0.918	0.93

Table 1: Results

To evaluate the accuracy of name generation, we analyzed inputs and outputs by hand and compared these manually defined groups with automatically created ones. First, all correct results were manually identified from the input data which gives us the number of *existing values*. The cluster was considered as valid if it did not contains false positives and there was no other cluster that contained tabular descriptions of the same instance. The number of correctly created clusters defines the number of *correct found values*. The number of *found values* corresponds to the size of the output. These values are then used to calculate standard information retrieval measures: *Precision*, *Recall* and *F-Measure*.

For the domain of digital cameras the system created 498 groups, i.e., 498 camera models have been located for which more than one description existed. 234 groups were generated for the domain of notebooks. The numbers of the final results are presented in Table 1. As we can see, the F-Measure for the digital camera domain is higher than for notebooks, since more observations were exploited. Note that these measures will improve if more observations are available. Thus, we view our results to be very promising as we could achieve very accurate cluster analysis although we only relied on a limited number of observations.

We plan to perform extensive evaluations on additional domains of consumer electronics like cell phones and MP3 players, as well as on other domains, in which instances are described in tabular form.

Acknowledgments

The research project is funded partly by the grant from the Austrian Research Promotion Agency, Programm Line FIT-IT Semantic Systems (www.fit-it.at), Project AllRight, Contract 809261.

4. REFERENCES

[1] D. Pelleg and A. W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Seventeenth International Conference on Machine Learning*, pages 727–734. Morgan Kaufmann, 2000.

[2] G. Stoilos, G. B. Stamou, and S. D. Kollias. A string metric for ontology alignment. In *International Semantic Web Conference*, pages 624–637, 2005.