Automatic Extraction of attribute information by using Anchor Text

Shigeru Ohta (510019)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 8, 2007

Keywords: anchor text, k-means, clustering, attribute, site map.

This paper proposes a method that acquires synonymous anchor texts (e.g., access, transportation) that correspond to attributes of a given class (e.g., univerisity) from web sites that describe objects (e.g., JAIST, Kanazawa univ.) in the same class.

Recently, people have generated numerous official homepages (web sites) for a wide range of objects in various classes such as hospital or university, and these web sites are regarded to be a principal source for users to obtain information on the objects.

However, even if we would like to examine and compare information on various objects in the same class, we often need more time to find the target informatoin than we had expected because the target information are frequently referred by different anchor text. For example, let us assume that a user would like to find informatoin on the address of several universities from their websites. Although the user need to follow a link to the document that include the target information, the document is often linked through different synonymous anchor texts such as "transportation," "access," and "map." We therefore need some time to match the anchor text

with the attribute we want to know, or even read through entire pages linked through putative anchor texts.

In this paper, we assume that when websites describe objects in the same class, those websites include pages that mention information on similar attributes in that class, and propose synonymous anchor texts that referred to pages that describe the same attributes. Here, when a document in a website for an object describe the information of the attribute of the object, anchor texts are usually means the attribute of that object. Thus, acquired synonymous anchor texts are also synonymous attributes of that object. If we can acquire synonymous anchor texts from websites for objects in the same class, we can construct the standard site map for that class. This standard site map includes the common attributes name as anchor texts. We can easily reach the target information by using a mapping between pages in a website for an object and the standard site map for the class of that object. Also, site designers can improve usability and accessibility of their websites for objects by designing their websites according to the standard site maps for the classes of the objects.

In this study, we acquire synonymous anchor texts from websites for objects in the same class in the following steps.

We first extract link pairs (a pair of an anchor text and its referring document) and select link pairs whose anchor texts are appropriate for attributes of the obejcts (Step 1). Then, we generate a term vector for each document in the link pairs (Step 2). Finally we perform clustering of those document by running k-means clustering on their term vectors. In the following, we describe each step in detail.

Step 1 We recursively retrieve pages in each website and extract link pairs (a pair of an anchor text and its referring document) and select link pairs whose anchor texts are appropriate for attributes of the obejcts. Link pairs are selected by judging whether anchor texts are appropriate as candidate of attribute names. Anchor texts are eliminated

1) when the length of the anchor texts exceeds a certain threshold or 2) when the anchor texts refers to pages in different websites, or 3) when the anchor texts includes stop words (e.g., “戻る (back),” “こちら (here),” “ジャンプ (jump).”)

Step 2 We then construct a term vector for each document in the link pairs. We only used terms extracted from each document by exisiting term extractor, and set their weight to their $TF - IDF$ values in that document. When the terms are included in title, meta, or heading tags in that document, or included in anchor texts that referred to the document, we gave some bonus to their weights. Since some pages include enough terms to create an informative term vector, we also used urls of the pages and anchor texts that referred to the pages to generate term vectors.

Step 3 We finally perform k-means clustering for term vectors generated in Step 2, and extract synonymous anchor texts from the resulting clusters. The number k of the resulting clusters are simply set to the number of pages in the web sites that included the least number of pages among all the websites used in the experiments.

We conducted experiments on each five websites for objects in two classes (laboratory and hospital). Although resulting clusters are rather noisy in the sense that one cluster often include anchor textsthat referred to information on different attributes of the objects,synonymous anchor texts such as “構成員一覧” and “メンバー” for an attribute “member” of a laboratory and “受診と入院の案内,” “入院のご案内,” “入院案内,” “入院案内, ” “入院案内” for an attribute “hospital admission info.”.