FINDING HIGH-QUALITY WEB PAGES USING COHESIVENESS

(Research-in-Progress)

Joshua C.C. Pun and Frederick H. Lochovsky
Department of Computer Science, Hong Kong University of Science & Technology, Hong Kong
{punjcc, fred}@cs.ust.hk

Abstract: For a document, cohesiveness is a measure of how closely the concepts in it are related to each other. Previous studies in linguistics have shown that a high quality document is likely to be very cohesive. Similarly, since a web page is a type of document, a high quality web page is expected to be very cohesive as well. Using an ontology constructed from the Yahoo! directory and the web pages linked to it as an underlying reference, we define a distance metric to measure how close two nodes (or concepts) are in the ontology. This metric is used to calculate the cohesiveness of a web page as the total distances of all the concepts in it. Users can use web page cohesiveness to more easily find high quality (cohesive) web pages.

Key Words: Cohesiveness, Web Quality

1. INTRODUCTION

The advent of the Web has created an information explosion. Billions of web pages are easily and freely available. However, their quality is often questionable as there are no controls on the web publication process. In most cases, there are no editorial reviewers to ensure the quality of a web page's contents. Hence, the ability to effectively find high quality pages is essential for web users. In previous work, we identified various aspects of a web page that can affect its quality. Moreover, a framework of web data quality dimensions was proposed (Figure 1) and the measurement of the appropriateness dimension and its use to find an appropriate page for users was investigated [12]. This paper focuses on the cohesiveness web data quality dimension, a facet of the usefulness dimension.

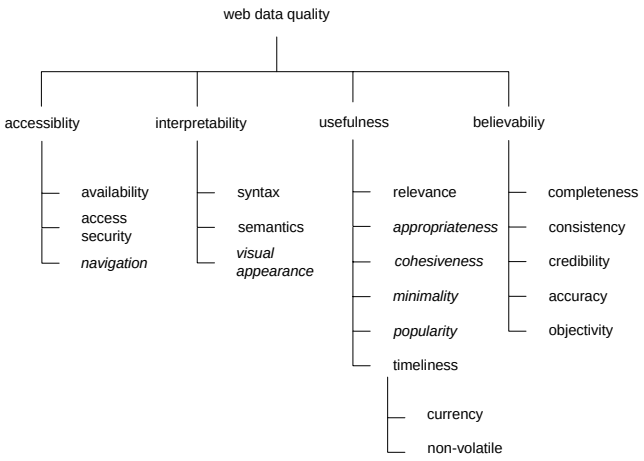

Figure 1: A Framework of Web Data Quality.

Recent research in psychology and linguistics has shown that an important factor in text comprehension is cohesion [9,11]. It is claimed that a text document with good cohesion (and coherence) can lead to easier understanding, which is an essential element of high quality information. By extension, if a web page is cohesive, it can be considered to be of high quality as well. Usually, to satisfy some information need, users would prefer to read a more cohesive web page since it would be more focused on the topic and thus be more useful to them. Currently, if two web pages are equally relevant to a user's query, search engines do not distinguish between the one that is more cohesive and the one that is less cohesive.

Our research addresses the issues related to identifying the cohesiveness features of a web page and measuring them automatically. By identifying the concepts (topics) in a web page, we develop a metric to calculate the cohesion among these concepts with reference to an ontology. The more related (tighter) the concepts are, the higher is the cohesiveness value of the web page and, thus, the higher the quality of the web page. If desired by the user, the cohesiveness values of a set of web pages can be used to identify and select those pages that meet the user's cohesiveness requirements.

The rest of this paper is organized as follows. The next section provides background information related to cohesiveness in the context of web pages. Its definition, its previous use in research on text documents and the available features in web pages are covered. Section 3 describes the steps used to measure the cohesiveness of web pages, starting from ontology construction to concept identification and representation. Section 4 introduces the metric for the cohesiveness dimension, presents an algorithm for measuring cohesiveness and illustrates, with examples, how the algorithm works. The evaluation of the metric is discussed in Section 5. Lastly, conclusions and future work are presented in the final section.

2. COHESIVENESS

2.1 Web Page Quality

Apart from our definition and framework of web page quality discussed in [12], previous research on the quality of web pages has focused mainly on the popularity and/or reputation of web pages, which is considered as being indicative of the quality of a web page [1,3,4]. Cho et al. define page quality as the probability of link creation by a new visitor, which is basically quite similar to PageRank [4]. They argue that the ranking of a search engine (e.g., Google using PageRank) is biased against new pages, which they term as the "rich-get-richer" phenomenon. Consequently, a new page takes some time to gain "popularity" even though it may, in fact, be a good quality page. Their work identifies these good quality pages early on. Similar to other previous work on page quality, their work only considers the popularity dimension within our web data quality dimension framework.

2.2 Definition of Cohesiveness

Depending on the context, there are many definitions of cohesiveness. For a web page, we define *cohesiveness* as a measure of how closely the *concepts* in it are related to each other. A cohesive page can help a user discover information much faster since it has fewer distractions from the main topic of interest to the user. The cohesiveness dimension considers all concepts that are contained in a web page. If the concepts in a web page are highly diversified and unrelated, the cohesiveness of the page is weak. If not, its cohesiveness is considered to be strong. This idea is, in fact, analogous to the concept of good programming practice in software engineering where the goal is to write a cohesive software module. In software engineering, *cohesion* is a measure of the degree to which a software component implements a single, focused function. A module is most cohesive when it does only one thing.

Cohesion for sequential text (or a document) is also well defined in computational linguistics [5,7]. It is the degree to which components of the text are linked. It is a characteristic of the text and is an objective property of the explicit language and text. In a document, there are explicit features, such as words, phrases, or sentences, that guide the reader in interpreting the ideas in the text, in connecting ideas with other ideas, and in connecting ideas to higher level global units (e.g., topics and themes). Computational linguistics also uses the notion of the *coherence* of a document, which is the extent to which a reader develops a unified situation model of the text. The coherence relations are constructed in the reader's mind and depend on the skills and knowledge of the reader. If the reader has adequate knowledge about the subject and there are adequate linguistic and discourse cues, then the reader can more easily form a coherent mental representation of the text. A reader perceives a text to be coherent if the ideas conveyed in the text are meaningful and presented in an organized manner. In short, coherence is a characteristic of the reader's mental representation of the text content.

Cohesion is a guide to coherence. An appropriate use of cohesive markers (or devices) can guide the reader to form a coherent representation of the document. Examples of cohesive markers are reference items (e.g., 'he', 'she', 'it') and conjunctions (e.g., 'and', 'but'). However, cohesive markers alone do not necessarily make a text coherent. Other plausible means to facilitate coherence include reader expectations of finding coherence and the framework provided by genre expectations (such as predictability inferred from setting and time sequence in narratives) and the structure of ordinary conversation (such as the adjacency pairs, question-answer and request-compliance) [2].

2.3 Measuring Cohesion in Text Documents

As text cohesion is a very crucial factor in comprehension, a computer tool to measure cohesion would definitely be beneficial for the evaluation of the quality of a text. Graesser et al. [7] have developed such a tool called Coh-Metrix. It measures "microscopically" every aspect of a text document, for example, word frequency, part of speech, densities of pronouns, logical operators and connectives, mean values of polysemy and hypernym, readability and so on. Coh-Metrix automatically evaluates text so as to eventually map the cohesion of the text to the background knowledge and reading skills of the reader. It also gives feedback to a writer about which aspects of the text are cohesive and which lack cohesion. This enables the writer to determine what aspects of the text need to be improved.

According to the current theories of discourse, a text has at least three levels of structure: the *surface code*, which refers to literal words used in the text, the *text-base*, which refers to the propositions that the surface code describes and the *situation model*, which refers to the representation of the world that the text is intended to convey. The situation model represents the world according to several dimensions, namely, causal, intentional, temporal, referential, spatial, and structural each of which has its own level of cohesion. Coh-Metrix estimates the cohesion of the text for each of these dimensions under the assumption that cohesion can be measured separately for each dimension.

2.4 Special Nature of Web pages

At first glance, one would think that measuring the cohesiveness of a web page should be quite similar to that of a text document. However, there are several fundamental differences between a web page and a text document [14]:

1. Too many distractions: The contents of a web page provide many more opportunities for distraction than a text document. For example, web pages contain hyperlinks (e.g., navigational links and advertisements) and make more frequent use of visual elements (e.g., images, colours and fonts). They can also be split into multiple blocks where each block is an independent information unit (or a document). These all can distract a reader's attention from reading a page continuously (even within the current page).

2. Discontinuous text processing: Readers of a web page have to choose which of the hyperlinks to follow after finishing the current page. Besides distracting their attention for text comprehension, as they have to be aware of "where they are" now, they also have to understand the relationship between the current page and the page to which they are going to click. As such, authors cannot always ensure the final topic continuity the readers will perceive.

3. No fixed reading sequence: A sequential text document, published in printed media, presents the content in a fixed sequence. This sequence will facilitate the author to build coherence for the readers. However, such a sequence often does not exist in web pages. Consequently, authors have many uncertainties about how people read their web pages. For example, have readers read the pages that (logically) precede this page? As a result, a web page (e.g., a subtopic of a page) has to be written in a way that it can be read in various, unpredictable sequences.

4. Vague document boundaries: Unlike printed documents, web pages do not have fixed and well-defined boundaries. Readers often cannot easily estimate the size and structure of the web document and they may be unaware of crossing the boundary of a web document when surfing. Again, this makes it more difficult for authors to build coherence than in a sequential text.

All of these factors make it difficult to directly apply the techniques and findings from Graesser's study [7] on the cohesion analysis of text documents to web pages.

2.5 Levels of Features in Web Pages

Depending on the task and its purpose, different features from a web page can be used. These features can be classified into at least three levels. First, the *surface* level includes those features related to the basic counts (e.g., number of words, tokens, sentences and visual elements) in a web page. This level of features was measured in the metrics developed for the *appropriateness* web data quality dimension [12]. The *linguistic* level involves a deeper level of features since it includes a detailed analysis of the linguistic properties of each word and sentence in a text including senses, types, imageability, clarity and so on. With these features, the comprehensibility of a document can be analyzed [7]. Finally, an even deeper level of features, the *semantic* level, is available. It considers the relationship of the concepts in a web page. To measure the cohesiveness of a web page, the features in this level are appropriate as we are interested in only a broader view of cohesiveness.

3. COHESIVENESS FOR WEB PAGES

3.1 Overview of Cohesiveness Measurement

A basic premise of our work is that whether two *concepts* (or topics) are related can be determined by the *closeness* (or *proximity*) of these two concepts in an ontology. For example, suppose we have three web pages (W_1, W_2 and W_3). W_1 contains only the concept *A*. W_2 contains two concepts *A* and *B* and these two concepts are quite related. Lastly, W_3 contains concepts *A* and *C*, but concept *A* and concept *C* are not related. Then, intuitively we would order the cohesiveness among these three web pages as $W_1 > W_2 > W_3$. If two web pages are equally relevant to a user query, users would probably prefer to read a more cohesive web page since it is more focused on a topic.

Before we can measure the cohesiveness of a web page, a concept hierarchy, represented as an ontology and used as a basis for comparing the similarities between concepts, first has to be constructed[1] (see Figure 2). To avoid the adverse effects of the kinds of distractions mentioned in Section 2.4 on the

[1] This step only needs to be done once as the ontology is not expected to change frequently.

determination of the cohesiveness of a web page, the data rich section[2] of a web page [15] has to be extracted and the concepts present in the data rich section need to be identified. Finally, knowing the concepts present in a web page and using the concept hierarchy, the cohesiveness of the concepts can be measured. In the following subsections, these steps will be further elaborated and explained.

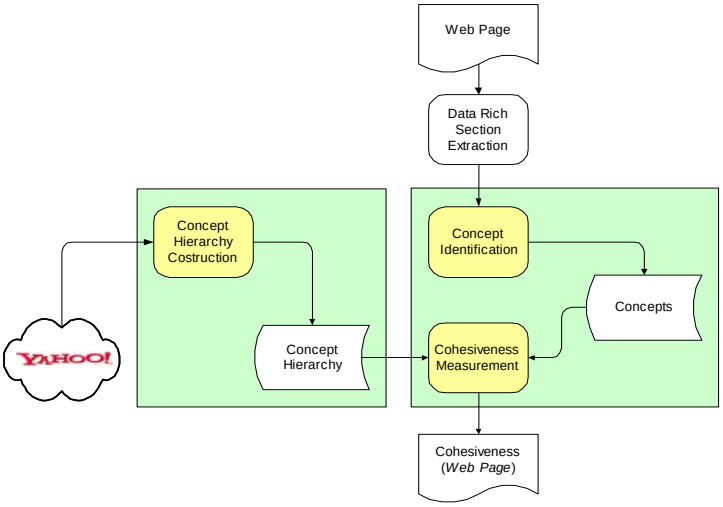

Figure 2: Overview of Cohesiveness Measurement.

3.2 Construction of Concept Hierarchy

3.2.1 Use and Extension of Ontology to Represent the Concept Hierarchy

To quantify how close two concepts are, an underlying reference is needed to make the comparison. An ontology is used as this reference where the nodes of the ontology represent the concepts. In addition to this underlying reference, a *distance* metric is proposed to measure how close two nodes (or concepts) are in the ontology. With this metric, the cohesiveness of a page can be calculated as the total distances of all concepts in it. This is an extension to the traditional use of an ontology for representing general and specific knowledge. This approach has also been used for scoring alternative speech recognition hypotheses [6] and for information retrieval in the web context [16].

3.2.2 Automatic Construction of Ontology from the Yahoo! Directory

While it is possible to manually construct an ontology for a particular domain, it is both time-consuming and inflexible to do so. In addition, the ontology should be updated constantly to reflect the ever-changing, dynamic real-world environment. This makes the manual approach not very feasible and an automatic approach is preferred. While there are some publicly available ontologies (e.g., WordNet), most of them are not suitable for our purpose. In WordNet, the focus is primarily on representing how words are related in terms of their parts of speech. In contrast, we want to represent real-world concepts (not just words) and relationships that are both general and specific since the ontology is used as a reference for comparing how close two concepts are. Hence, we use the Internet as a resource to create an ontology automatically. By using a well-maintained directory in the Internet[3] (i.e., the Yahoo! directory), an ontology can be constructed using the inputs from this directory and the web pages linked by them. Huang *et al.* have also used the Yahoo! directory as one of the web resources to build the corpus for their LiveClassifer prototype [8].

---

[2] The data rich section of a web page is the section or frame in its layout that contains the main content of the page.

[3] While we use the Yahoo! Directory in our work, any similar directory could be used.

The Yahoo! directory is manually maintained by Yahoo! and contains many categories. For each category, it is further classified into sub-categories, etc. In our case, the category name in the Yahoo! directory represents a *concept*. Figure 3 shows that the category *Computer Science* in the Yahoo! directory contains about 50 sub-categories. Following the category links in the Yahoo! directory to the last sub-category page (i.e., the directory page on which there are no further links to sub-categories), the category (or the concept) represented by this page will be determined by the text stored in the page. That is, the keywords on the page will be identified and used to represent this concept. For example, as shown in Figure 4, the text after the words "SITE LISTING" is used to describe the concept *Expert System*. Sometimes the text in the last sub-category page of the directory is too short (i.e., there are too few relevant words) to describe the concept. For example, there are less than 50 words in the web page of the *Expert System* concept. Using only a few words to describe a particular concept may not be very representative. To improve the representativeness, the text from the web pages linked to by the last sub-category page of the directory is also used to describe the concept. For the example in Figure 4, the three web links *"Java Expert System Shell"*, *"SHYSTER"* and *"FAQ – Expert System Shells"* will also be followed. The text from these three web pages, as well as the text from the original concept web page (i.e., the category *Expert System*), will be used to determine the best keywords to represent the concept.

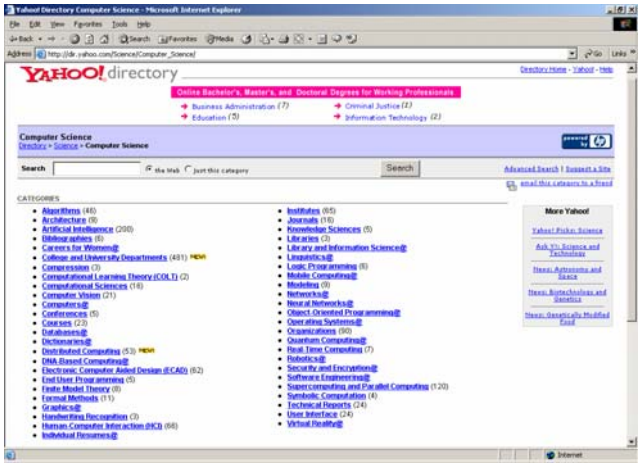

Figure 3: The category "Computer Science" in the Yahoo! directory.

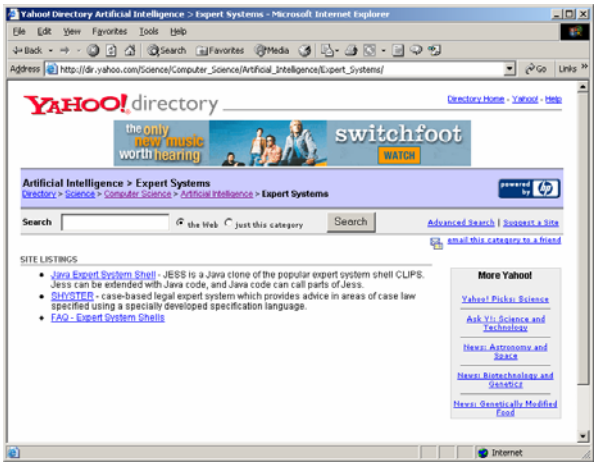

Figure 4: The category "Expert Systems" in the Yahoo! directory.

The hierarchy from the Yahoo! Directory, in fact, is a graph (not a tree). There are many cross-referenced concepts in the hierarchy which are represented with an "@" appended to the category name (see Figure 3). For example, the concepts *Networks, Neural Networks* and *Object-Oriented Programming* are cross-referenced concepts. For the development of our prototype, some restrictions were set on the construction of the concept hierarchy. The root of the concept hierarchy (or tree) that we use is only started from the concept *Computer Science* in the Yahoo! directory as this reduces the size of the hierarchy. Searching all the cross-referenced links in the Yahoo! directory will take a long time; hence an upper limit on the cross-referenced link level to be navigated and the maximum height of the concept hierarchy are set. This can avoid searching all the links endlessly as there may be loops in the Yahoo! directory[4].

[4] The upper limit for the cross-referenced link level to be navigated is determined by experimentally increasing it until there is no further improvement in the calculation of a web page's cohesiveness value. The maximum height is determined by a similar approach. Using this approach, the upper limit for the cross-referenced link level is set to 2 and the maximum height is set to 10.

3.3 Concept Representation and Identification

Many concepts can be contained in each web page. To identify the concepts in a web page, the formatting in the data rich section is first removed. The images, HTML tags and the hyperlinks are also ignored. This process reduces the content of the data rich section to plain text only. Next, some standard information retrieval pre-processing steps (e.g., stemming and stopword removable) are applied to this plain text.

The words (or terms) in the text with the highest term importance indicator are used to represent a concept. The term importance indicator (*tii*) is defined as:

$$tii_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log_2\left(\frac{N}{df_j}\right)$$

where

- tf_{ij} is the frequency of term j in document i
- df_j is the document frequency of term j (i.e., the number of documents containing term j)
- idf_j is the inverse document frequency of term j
- N is the number of documents in the collection

The term importance indictor tii_{ij} is a combination of both TF and IDF, which are commonly used in information retrieval, and is used to distinguish those words that can represent document i. A large value of tii_{ij} means that word j frequently appears in document i, but not in other documents. Hence, word j should be used to represent document i. In our case, a list of words, which have the highest term importance indicator, are used to represent a concept.

4. COHESIVENESS METRIC

4.1 Formulas

Intuitively, to determine the cohesiveness of a web page we need to measure how interrelated (or cohesive) are the concepts in the web page. Using the underlying reference ontology, the concepts in a web page can be grouped together to form concept clusters. A concept cluster is a group of nodes in an ontology all of which are related to each other by parent-child relationships. A distance metric also needs to be defined to measure how close two nodes (or concepts) are in the ontology. Intuitively, the cohesiveness of a web page depends on how cohesive and frequent are the concept clusters in a web page (i.e., how "close together" in the reference ontology are all the concept clusters in a web page and how often do they appear in a web page). We refer to this as *inter-cluster cohesiveness*. Inter-cluster cohesiveness, in turn, depends on how cohesive are the concepts in each concept cluster (i.e., how "close together" in the reference ontology are all the concepts in a concept cluster). We refer to this as *intra-cluster cohesiveness*.

The formula to measure the inter-cluster cohesiveness of a web page *intercoh(WP)* is:

$$intercoh(WP) = L(WP) \times \sum_{i=1}^{N} freq_i \times intracoh(C_i)$$

$$freq_i = \frac{A(C_i)^2}{\sum_i A(C_i)^2}$$

$$L(WP) = e^{\min(\frac{word(WP)}{100} - 1, 0)}$$

where

- N is the number of clusters in web page *WP*
- $L(WP)$ is the length factor of web page *WP*
- $intracoh(C_i)$ is the intra-cluster cohesiveness of cluster C_i
- $A(C_i)$ = the appearance frequency of a cluster in a web page calculated as the total number of times the concepts in C_i appear in web page *WP* (since a concept can appear more than once in a web page)
- $freq_i$ = the relative frequency of concepts in C_i appearing in web page *WP*

The above formula defines the inter-cluster cohesiveness of a web page as the sum of the intra-cluster cohesiveness of its concept clusters ($intracoh(C_i)$) multiplied by a factor to account for a cluster's relative frequency ($freq_i$). Based on a common recommendation for readability metrics[5], a length factor $L(WP)$ is used to adjust the cohesiveness value of web pages when they are shorter than 100 words.

A goal of the above formula should be to "*always encourage a large cohesive concept cluster, but penalize small, scattered and isolated concept clusters*". Furthermore, if a concept cluster appears very often in a web page, there should be a direct positive effect on the cluster's cohesiveness and ultimately on the cohesiveness of a web page. To "encourage" large cohesive clusters that appear often in a web page, the formula is biased towards always awarding a larger cohesiveness value to bigger clusters and a smaller cohesiveness value to scattered, isolated and independent clusters by means of the relative frequency, $freq_i$. The relative frequency is calculated as the square of the appearance frequency, $A(C_i)$, of a cluster in a web page divided by the sum of the squares of the appearance frequency of all clusters. The use of the square function further intensifies the effect of awarding a large value to bigger clusters and a small value to scattered, isolated and independent clusters.

The formula to measure the intra-cluster cohesiveness of each concept cluster C_i is:

$$intracoh(C_i) = w_i \times \frac{\sum_{\forall x,y \in C_i, x \neq y} close(x,y)}{\max(N_c(C_i) - 1, 1)}$$

$$w_i = \frac{N_c(C_i)^2}{\sum_i N_c(C_i)^2}$$

$$close(x,y) = \frac{overlap(x,y)}{overlap(x,y) + height(z,x) + height(z,y)}$$

where

- $N_c(C_i)$ is the number of nodes (or concepts) in cluster C_i
- w_i is the weight factor for cluster C_i
- `close(x,y)` is the closeness of two nodes, `x` and `y`
- `overlap(x,y)` is the number of common edges in the path from the `Root` to node `x` and node `y` (i.e., the height from `Root` to node `z` as shown in Figure 5)
- node `z` is the **nearest common ancestor** of node `x` and node `y`
- `height(z,x)` and `height(z,y)` are the height from node `z` to node `x` and node `y`, respectively

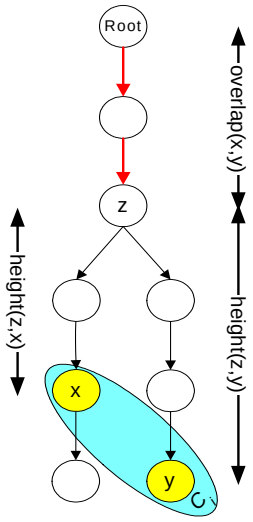

Figure 5: Close Function.

[5] [13] recommends that readability metrics be used only on texts of at least one hundred words to get meaningful results.

The intra-cluster cohesiveness of a concept cluster, *intracoh(C_i)*, is the product of its weight factor (w_i) and its average closeness when pairing up nodes in the cluster. The weight factor (w_i) is calculated as the square of the number of nodes in cluster C_i divided by the sum of the squares of the number of nodes in all clusters. The use of the square function intensifies the effect of awarding a large value to bigger clusters and a small value to scattered, isolated and independent clusters. The average closeness is the sum of the closeness of the nodes in a concept cluster, $\sum$close(x,y), divided by the total number of nodes paired up in the cluster (i.e., $N_c(C_i) - 1$). To avoid division by zero, this total number is set to at least 1 (i.e., we take max($N_c(C_i) - 1$, 1)). Intuitively, the closeness value of two nodes in a concept cluster is related to two issues:

- The length of the common path (i.e., path overlap) in the ontology from the root to two paired up nodes (i.e., the longer the common path the closer are the two concepts and, thus, the larger is the closeness value).
- The level in the ontology of the two paired up nodes (i.e., the farther apart in the level of the hierarchy the paired up nodes are, the farther apart are the two concepts and, thus, the smaller is the closeness value).

As stated above, the cohesiveness of a cluster depends on the average closeness of the nodes in it. To determine the closeness of nodes, a special algorithm has been devised. Instead of selecting nodes from a cluster randomly to calculate their closeness, two nodes are chosen each time in such a way that they give the largest value of closeness when paired up. To achieve this, a node always pairs up with its immediate parent, if possible. Otherwise, it pairs up with its siblings (or children of siblings). This pairing up process is done once for each node until all nodes in a cluster have been considered. Calculating the cohesiveness value of a cluster in this way can ensure that the largest possible cohesiveness value is achieved. Otherwise, the cohesiveness value of a cluster may be underestimated if the nodes are selected randomly so that even a cohesive cluster would not have a high cohesiveness value. Hence, the closeness function, *close(x,y)*, involves three values: *overlap(x,y)*, *height(z,x)* and *height(z,y)*. The height values are in fact inversely related to the level of the nodes in the hierarchy. Furthermore, the higher is the level of the two nodes in the hierarchy, the larger is their closeness value.

4.2 COHEM algorithm

Figure 6 presents the algorithm for determining the cohesiveness measurement (COHEM algorithm). In the algorithm, issues related to the calculation of inter- and intra-cluster cohesiveness are refined in steps. Intuitively, the algorithm is composed of three main steps:

i. Form concept clusters of the web page according to the concept hierarchy.
ii. Calculate the intra-cluster cohesiveness of each concept cluster.
iii. Calculate the inter-cluster cohesiveness of a web page by combining the intra-cluster cohesiveness of its concept clusters.

```
(* Form concept clusters. *)
for all nodes j ∈ C_WP
   for concept cluster C_i ∈ all concept clusters
      if   Family(j) ∧ Family(C_i) ≠ ∅
     then  insert node j to C_i
     else  form new cluster containing node j
     endif

(* Calculate intra-cluster cohesiveness for each concept cluster.                *)
for C_i ∈ all concept clusters
   for all nodes j ∈ C_i (* breadth-first search *)
      add node j to WorkSet_i
      if   only ONE node in C_i
      then
        intracoh(C_i) = close(j, j) = 1
```

```
     else
       (* For multi-node clusters, find node x to pair up with node j. To calculate *)
       (* their closeness, x is a node ∈ WorkSet_i-j s.t. common path of node x &    *)
       (* node j to Root is longest.                                                 *)
       if WorkSet_i-j ≠ ∅
       (* Ignore 1st node ∈ multi-node cluster. *)
       then
          x = LongestCommonPath(WorkSet_i-j, j)
          calculate close(x, j)
          add close(x, j) to intracoh(C_i)
       endif
     endif
   intracoh(C_i) = w_i * (intracoh(C_i) / (N_c(C_i)-1))

(* Calculate the inter-cluster cohesiveness of the web page by combining the *)
(* intra-cluster cohesiveness values of its concept clusters.                *)
for C_i in all concept clusters
   add intracoh(C_i)*freq(C_i) to intercoh(WP)
```

Figure 6: COHEM algorithm.

where

- C_{WP} are the concepts appearing in web page WP
- Family(j) is the set of family nodes of node j (including ancestor nodes of node j and node j itself, but excluding the Root node). An ancestor node of any node j is any node above j in a tree model where "above" means "toward the root". In Figure 7, Family(B)={B}, Family(N)={B,H,N}.
- Family(C_i) is the set of family nodes from all nodes in concept cluster i. That is, $\text{Family}(C_i) = \bigcup_{\forall j \in C_i} Family(j)$
- intracoh(C_i) is the sum of close value of C_i.
- WorkSet_i is a temporary set variable containing nodes, which have been considered in C_i so far.
- LongestCommonPath(WorkSet_i-j,j) returns a node, which is in the set WorkSet_i-j and gives the longest common path from Root to node j and from Root to the returned node.

4.3 Illustration of COHEM algorithm

The illustration is based on the concept hierarchy in Figure 7. There are four concept clusters (or clusters) and each cluster has one or more concepts (or nodes). The steps in the algorithm are:

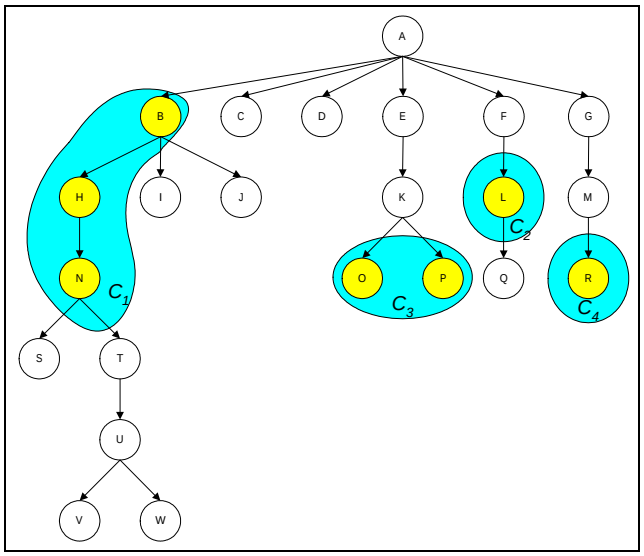

Figure 7: Web page with 7 concepts in 4 clusters.

i. Find the concepts (or nodes) contained in a web page, WP. (i.e., C_{WP} = {B,H,N,O,P,L,R}).

ii. Based on the COHEM algorithm, form four concept clusters (C_i) :

- C_1 = {B,H,N} $N_c(C_1) = 3$
- C_2 = {L} $N_c(C_2) = 1$
- C_3 = {O,P} $N_c(C_3) = 2$
- C_4 = {R} $N_c(C_4) = 1$
- $\sum_{i}^{4} N_c(C_i)^2 = 3^2 + 1^2 + 2^2 + 1^2 = 15$

iii. Calculate the intra-cluster cohesiveness of each concept cluster C_i.

Cluster	Cohesiveness of cluster
C_1	$close(B,H) = \frac{1}{1+0+1} = \frac{1}{2}$ $close(H,N) = \frac{2}{2+0+1} = \frac{2}{3}$ $intracoh(C_1) = \frac{9}{15} \times \left[\frac{close(B,H)+close(H,N)}{2}\right] = \frac{9}{15} \times \left[\frac{\frac{1}{2}+\frac{2}{3}}{2}\right] = \frac{7}{20}$
C_2	$close(L, L) = 1$ $intracoh(C_2) = \frac{1}{15} * close(L, L) = \frac{1}{15}$
C_3	$close(O,P) = \frac{2}{2+1+1} = \frac{2}{4} = \frac{1}{2}$ $intracoh(C_3) = \frac{4}{15} * close(O,P) = \frac{4}{15} \times \frac{1}{2} = \frac{2}{15}$
C_4	$close(R, R) = 1$ $intracoh(C_4) = \frac{1}{15} * close(R, R) = \frac{1}{15}$

iv. Derive the inter-cluster cohesiveness of a web page by summing up the intra-cluster cohesiveness of each concept cluster multiplied by its appearance frequency. In this example, all concepts are assumed to appear once in the web page.

$$intercoh(WP) = \sum_{i=1}^{4} freq_i \times intracoh(C_i)$$

$$intercoh(WP) = \frac{9}{15} \times \frac{7}{20} + \frac{1}{15} \times \frac{1}{15} + \frac{4}{15} \times \frac{2}{15} + \frac{1}{15} \times \frac{1}{15} = 0.2544$$

5. EFFECTIVENESS OF COHESIVENESS METRIC

5.1 Background

Two evaluations were designed to verify that the cohesiveness metric measured by the COHEM algorithm is comparable to a user's judgment regarding a web page's cohesiveness. To facilitate the discussion, we define the following notation:

- $W = \{w_1, \ldots, w_i, \ldots, w_N\}$ where N is the number of web pages.
- $S = \{s_1, \ldots, s_j, \ldots, s_M\}$ where M is the number of subjects.

The cohesiveness of each web page w_i, $cohesiveness(w_i)$ is evaluated both by the COHEM algorithm (COHEM) and by the subjects and is represented as follows:

$w_i \rightarrow COHEM\ algorithm \rightarrow cohesiveness\ (w_i, COHEM)$
$w_i \rightarrow Judgment\ (s_j) \rightarrow cohesiveness\ (w_i, s_j)$

The objective is to verify whether: $\forall i, j$, **cohesiveness(w_i, COHEM) ≈ cohesiveness (w_i, s_j)**. In making this comparison, several issues need to be addressed such as whether $\forall j$ *cohesiveness* (w_i, s_j) is consistent.

Our evaluations involved seven subjects. Their academic backgrounds varied from associate degree level to graduate degree level. Their ages ranged from 24 to 34, while their Internet experience ranged from 4 to 8 years.

5.2 Design of Initial Evaluation

It is difficult to request an absolute value of *cohesiveness* (w_i, s_j) from subject s_j for web page w_i. At the same time, it is also very difficult to ask subjects to order *N* web pages (if *N* is large) by cohesiveness as the cohesiveness difference between any two web pages may not be that clear. To resolve this issue, instead of requesting an exact cohesiveness value for each web page from subjects, it is more feasible to ask subjects to categorize web pages into a few broad, discrete cohesiveness categories. Indeed, a rough relative ordering on the cohesiveness of web pages is more important than determining their absolute values. Hence, subjects were requested to classify *N* web pages into three different categories:

- Strongly cohesive (SC)
- Cohesive (CO)
- Weakly cohesive (WC)

Our test set contained 91 web pages. Their cohesiveness values were measured by the COHEM algorithm and they were also classified by subjects according to the above categories. Based on the subjects' judgments, the distribution of the cohesiveness values for each category is shown in the table and graph in Figure 8. In the graph in Figure 8, the taller the "hill", the greater the number of web pages for that

cohesiveness category. The graph shows that there are more "Cohesive" web pages and fewer "Weakly Cohesive" and "Strongly Cohesive" web pages. As the variation in the cohesiveness value of each category is not large and the medians of the cohesiveness value of each category are quite far apart, the cohesiveness metric of a web page is a good estimate of a subject's judgment.

	Range of cohesiveness value		
Category	***Lowest***	***Highest***	***Median***
SC	0.2462	0.4084	0.3458
CO	0.1223	0.5045	0.1863
WC	0	0.1615	0.1457

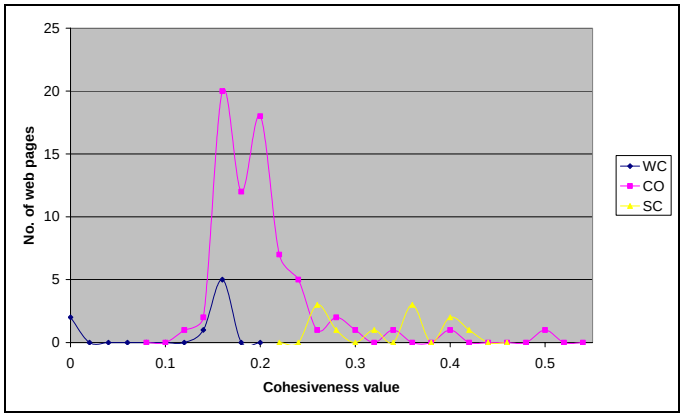

Figure 8: Range of cohesiveness value for different cohesiveness categories.

5.3 Design of Second Evaluation

In the initial evaluation, the subjects indicated that occasionally they had difficulty in grouping some of the web pages into the three categories. The goal of this second evaluation was to further simplify the decision made by the subjects. In this evaluation, subjects only needed to decide whether the current test page was more or less cohesive than a standard reference set.

The design of the evaluation is as follows. First, by using the COHEM algorithm, the test web page set (W) is ordered by its cohesiveness value and a few web pages with "middle cohesiveness" are selected from the set W as the *reference set* (Φ). Next, a web page W_p is randomly picked from the web page set (W-Φ). The subject is then asked to compare the current page W_p with the reference set (Φ) and decide, in terms of cohesiveness, whether W_p is more, equally or less cohesive. We repeat this process for N pages.

If a subject classifies W_p as more cohesive and if the cohesiveness value of W_p is greater than that of Φ, then the cohesiveness metric gives a correct estimate. If the majority of the classifications from the subjects conform to the classification given by the cohesiveness metric, then this shows that the cohesiveness metric is a good estimate for the cohesiveness of a web page.

In the evaluation, three web pages (i.e., those ranked 45^{th} to 47^{th} in terms of cohesiveness) in the test set were defined as the reference set. Their cohesiveness value ranged from 0.18854 to 0.19213. Seven users were briefed regarding the definition of the cohesiveness of a web page and then requested to compare the cohesiveness of a set of web pages with the standard reference set as described above. Each user had to compare 20 web pages with the reference set and decide whether they were more, equally or less cohesive than the reference set. Excluding the best and worst performers in the evaluation, the classifications from the middle five users are used. 74% of the classifications from these users conformed to the classification given by the cohesiveness metric. If an error allowance of 10% is allowed on the cohesiveness value, the percentage of correct classifications increases to 82%.

On the issue of whether there is consistency among the judgments from users on the cohesiveness of a web page, the results show that ~77% of the judgments agreed with each other[6]. Thus, we conclude that the judgment of subjects regarding the cohesiveness of a web page is quite consistent. This is important as it shows that there is some degree of objectiveness to this measure even though it is a quite subjective issue in itself.

6. CONCLUSIONS

The quality of web pages is always in doubt in the current web environment as there are no controls on the publication process. Since most web users prefer to find high quality web pages, an effective means to identify high quality pages is needed. Search engines normally use the web data quality dimensions of relevance, believability and popularity as their primary measures of quality when finding and ranking web pages. Sometimes, other dimensions of web page quality are also important to users and should be considered. One such dimension of web page quality is *cohesiveness*. The cohesiveness dimension gauges how closely the concepts are related to each other in a web page. The contribution of this paper is to propose a metric to measure this dimension using the semantic level of web page features and to develop an algorithm to determine it automatically.

By using a well-maintained directory in the Internet (i.e., Yahoo! directory), an ontology is constructed using the inputs from the directory and the web pages linked by it. This ontology is not just used for representing the knowledge structure, but also for estimating the cohesiveness of concepts (or nodes) within it. On top of this underlying reference, a distance metric is defined to measure how close two nodes are in the ontology. The cohesiveness of a page is then calculated as the total distances of all concepts in it.

Evaluation results have shown that the judgments on cohesiveness of web pages from users are fairly consistent. Moreover, the rating of the cohesiveness of a web page from subjects is comparable to the value given by the cohesiveness metric. Hence, it is possible, with the help of the cohesiveness metric, for users to select strongly cohesive pages more easily. This can eventually be useful for finding high quality pages.

Several issues regarding cohesiveness are worth further investigation. Generally, a strongly cohesive web page is preferred. One issue to investigate is whether there is any relationship between web genres and the cohesiveness of web pages. In text documents, it is claimed that a high cohesive level is not always desirable [11]. When the prior knowledge level of readers on a particular topic is high, a highly cohesive text can sometimes impair their comprehension. While this may appear to be counter-intuitive, the reason for this phenomenon is that if a text is highly cohesive, then knowledgeable readers only pay cursory attention to the content. On the other hand, if the text is less cohesive, then readers are forced to pay closer attention to the content, thus engaging them in compensatory processing at the level of the situation model, which enables them to understand the text more deeply than if a very cohesive text is given to them. Whether this phenomenon is also true for web pages would be interesting to investigate. Finally, larger scale experiments with more subjects can be conducted to further verify our results.

[6] The results also show that only 15% of web pages are consistently misclassified by at least two users' judgments.

ACKNOWLEDGEMENT

This research was supported by the UGC Research Grants Council of Hong Kong under grant HKUST 6172/04E.

REFERENCES

[1] B. Amento, L. Terveen and W. Hill. “Does “Authority” mean quality? Predicting expert quality ratings of web documents,” *Proc. 23rd ACM SIGIR Conf.*, 296-303, 2000.

[2] Jungok Bae. “Cohesion and Coherence in Children’s Written English: Immersion and English-only Classes”, *Issues in Applied Linguistics*, **12**(1), 51-88. 2001.

[3] R. Baeza-Yates, F. Saint-Jean and C. Castillo. “Web structure, dynamics and page quality,” *Proc. SPIRE 2002*, LNCS, Springer, 2002.

[4] Cho Junghoo, and Sourashis Roy "Impact of Search Engines on Page Popularity." In *Proceedings of the 13th International Conf. on World Wide Web*, 20-29, May 2004.

[5] Dufty, D. F., McNamara, D., Louwerse, M., Cai, Z. and Graesser, A. C. “Automated Evaluation of Aspects of Document Quality,” *Proceedings of the 22nd Annual International Conf. on Design of Communications*, 14-16, 2004.

[6] Iryna Gurevych, Rainer Malaka, Robert Porzel and Hans-Peter Zorn. “Semantic coherence scoring using an ontology”, Proceedings of the Joint Human Language Technology and Northern Chapter of the Association for Computational Linguistics Conference (HLT-NAACL), Edmonton, Canada, 88 – 95, 2003.

[7] Graesser, A., McNamara, D.S., Louwerse, M., and Cai, Z. “Coh-Metrix: Analysis of Text on Cohesion and Language”, Behavioral Research Methods, Instruments, and Computers **36**, 193-202, 2004.

[8] C-C Huang, S-L Chuang and L-F Chien. “LiveClassifier: Creating Hierarchical Text Classifiers through Web Corpora”, *Proc. of the 13th International Conference on World Wide Web*, 184-192, 2004.

[9] Irwin, Judith. “Cohesion and Comprehension: A Research Review.” *Understanding and Teaching Cohesion Comprehension*, Ed. Judith Irwin, International Reading Association, 31-43, 1986.

[10] B.J. Jansen, A. Spink and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web,” *Information Processing and Management* **36**(2), 207–227, 2000.

[11] McNamara, D.S., Kintsch, E., Songer, N.B., and Kintsch, W. “Are good texts always better? Text coherence, background knowledge and levels of understanding in learning from text.” *Cognition and Instruction* **14**, 1-43, 1996.

[12] Joshua C.C. Pun and Frederick H. Lochovsky, "Ranking Search Results by Web Quality Dimensions," *Journal of Web Engineering* **3** (3&4), 2004, 216-235.

[13] Readability.info. “*Readability Scores, Grades, Sentences, Paragraphs, Word Usage, English Usage.*” http://www.readability.info/info.shtml

[14] Storrer A. “Coherence in Text and Hypertext,” *Document Design* **3**(2), 156-168, 2002.

[15] Jiying Wang, and Fred Lochovsky. Data-rich Section Extraction from HTML pages, *Proc. of the 3rd International Conference on Web Information System Engineering (WISE2002)*, 313-322, Dec 2002.

[16] Zhu, Xiaolan, and Gauch, Susan. “Incorporating Quality Metrics in Centralize/Distributed Information Retrieval on the World Wide Web,” *Proc. of the 23rd International ACM SIGIR Conf.*, 288-295, 2000.