A Precise Metric for Measuring How Much Web Pages Change*

Shin Young Kwon[1], Sang Ho Lee[1], and Sung Jin Kim[2]

[1] School of Computing, Soongsil University,
1-1 Sangdo-dong Dongjak-gu, Seoul 156-743, Korea
{sykwon, shlee}@comp.ssu.ac.kr
[2] School of Computer Science and Engineering, Seoul National University,
San 56-1 Shinlim-dong Kwanak-gu, Seoul 151-744, Korea
sjkim@oopsla.snu.ac.kr

Abstract. A number of similarity metrics have been used to measure the degree of web page changes in the literature. When a web page changes, the metrics often represent the change differently. In this paper, we first define criteria for web page changes to evaluate the effectiveness of the metrics in terms of six important types of web page changes. Second, we propose a new similarity metric appropriate for measuring the degree of web page changes. Using real web pages and synthesized pages, we analyze the five existing metrics (i.e., the byte-wise comparison, the TF·IDF cosine distance, the word distance, the edit distance, and the shingling) and ours under the proposed criteria. The analysis result shows that our metric represents the changes more effectively than other metrics. We expect that our study can help users select an appropriate metric for particular web applications.

1 Introduction

In many web applications, administrators create and manage web databases (a collection of web pages). Major web search service providers such as Google or Yahoo create web databases, with which users can conduct search activities. Proxy servers and web browsers maintain web databases to cache web pages and reduce repeated downloading of the web pages. As web pages change dynamically, web databases become obsolete and need to be updated. Updating all the web pages in the databases often entails making unnecessary requests and downloading unchanged web pages. Administrators would like to update only changed (or significantly changed) web pages in the databases, hence it is important to know how much the contents of web pages changed.

A number of similarity metrics for textual data have been used to measure the degree of web page changes. The simplest way to see if a web page changes is to compare web pages in a byte-by-byte level, which is used in [1, 3, 7]. Ntoulas et al. [9] used the TF·IDF cosine distance and the word distance. Lim et al. [8] used a metric based on the edit distance. Broder et al. [2] and Fetterly et al. [6] used the shingling metric. Each of the metrics often represents the same change of web pages

* This work was supported by Korea Research Foundation Grant (KRF-2004-005-D00172).

M.L. Lee, K.L. Tan, and V. Wuwongse (Eds.): DASFAA 2006, LNCS 3882, pp. 557 – 571, 2006.

differently. Users may have a difficulty with selecting an appropriate metric for their specific applications. In our best knowledge, there have been no research activities to intensively compare (or evaluate) the existing metrics in terms of web page changes.

In this paper, we propose criteria for web page changes in order to evaluate existing similarity metrics. In the criteria, web page changes are classified into six types (namely, "add", "drop", "copy", "shrink", "replace", and "move"). We believe that the six types represent common changes on the web. For each of the six change types, each criterion is defined. Each criterion gives appropriate values of change degree for the web pages changed by each change type. A metric is defined to be effective if the metric is close to the criteria. A metric is defined to be oversensitive (undersensitive) if the metric is always higher (lower respectively) than the criteria. In this paper, we also present a new similarity metric measuring the degree of web page changes effectively. The metric is designed to reflect our criteria well in terms of the six change types.

We conducted two kinds of experiments. The first experiment shows how differently the five existing metrics (i.e., the byte-wise comparison, the TF·IDF cosine distance, the word distance, the edit distance, and the shingling metrics) and ours represent the same change of web pages with 41,469 real web pages. In the second experiment, we evaluate the effectiveness of the six metrics with synthesized pages under the criteria. From the results, we substantiate that the existing metrics have some drawbacks and our metric is more effective than the existing metrics for web page changes.

This paper is organized as follows. Section 2 explains the existing metrics briefly. In section 3, the change types of web pages are defined, and the criteria for web page changes are described. In section 4, we propose an effective similarity metric for measuring the degree of web page changes. Experimental results are reported in section 5. Finally, section 6 contains the closing remarks.

2 Existing Metrics

We introduce five metrics that have been used to measure the degree of web page changes in the literature. Throughout this paper, p denotes an original web page and p' a changed page of p. The byte-wise comparison metric [1, 3, 7], which compares two web pages p and p' sequentially character by character, is the simplest (but most rigid) method for measuring the degree of web page changes. The metric returns 0 when there is no change at all, and returns 1 otherwise. The byte-wise comparison metric returns 1 even for very trivial changes (for example, insertion of one blank space). That metric is oversensitive and does not represent the degree of changes.

The TF·IDF cosine distance metric is commonly used for determining relevance of documents to a search query in the field of information retrieval. This metric transforms p and p' to the TF·IDF weighted vectors v_p and $v_{p'}$ respectively [10], and calculates cosine distance between the two vectors as equation (1). $v_p \cdot v_{p'}$ denotes the inner product of v_p and $v_{p'}$, and $||v_i||_2$ denotes the second norm of vector v_i.

$$D_{COS}(p, p') = 1 - \frac{v_p \cdot v_{p'}}{\|v_p\|_2 \|v_{p'}\|_2} \tag{1}$$

The word distance metric calculates how many of the words on a page have changed. In this metric, the distance between p and p' is calculated as equation (2), where m and n denote the numbers of words on p and p' respectively. Ntoulas et al. [9] used the TF·IDF cosine distance and the word distance to measure the degree of web page changes.

$$D_{WD}(p, p') = 1 - \frac{2 \cdot |common\ words|}{m + n} \tag{2}$$

The TF·IDF cosine distance and the word distance metrics cannot consider the change of word orders because they regard web pages as bags of words. The change of word orders frequently takes place and may be critical. For example, a change of word orders in a shopping site could represent a change of articles' priorities, which is important to customers.

The edit distance is the least expensive cost for sequences of edit operations (generally, "insertion", "deletion", and "substitution") required to transform one sequence to another sequence [5]. For example, suppose that the cost of every edit operation is 1. The edit distance from a sequence <A, G, B, A, A> to <A, B, A, T, A> is 2 because at least two edit operations (one "deletion" of G and one "insertion" of T) are needed. To measure the degree of web page changes, Lim et al. [8] defined a distance metric as equation (3), where m and n denote the numbers of words on p and p' respectively and δ denotes the edit distance between p and p'. Each page is regarded as a word sequence. Only two operations (i.e., "insertion" and "deletion") are used as edit operations, and the cost of the operations is 1. When the two pages are identical, δ becomes 0 and the metric returns 0. On the other hand, when the two pages are completely different, δ is $(m + n)$ because m old words are deleted and n new words are inserted. In this case, the metric returns 1.

$$D_{ED}(p, p') = \frac{\delta}{m + n} \tag{3}$$

The word distance and the edit distance metrics cannot distinguish insertion (deletion) of unique words from insertion (deletion, respectively) of duplicate words, even though such changes could have different implications. For example, if a web page $<w_1, w_2, w_3, w_4>$ changes to $<w_1, w_2, w_3, w_4, w_2, w_3>$, both metrics return 0.2. If the same page changes to $<w_1, w_2, w_3, w_4, w_5, w_6>$, they also return 0.2. However, it may be necessary to distinguish the two changes, because some applications like to consider them differently.

In the shingling metric, each web page is represented as a set of k-word continuous, ordered subsequences. Each subsequence is called a "shingle", and k represents the number of words on a shingle. Every word on the document starts a shingle wrapping at the end of the document. For example, the 3-shingling of a document $<w_1, w_2, w_3, w_4>$ is the set $\{<w_1, w_2, w_3>, <w_2, w_3, w_4>, <w_3, w_4, w_1>, <w_4, w_1, w_2>\}$. For a given shingle size, the distance between p and p' is shown in equation (4), where $S(p)$ is the set of shingles on p, and $|S(p)|$ is the number of elements on the set $S(p)$. Broder et al. [2] defined the shingling metric and tried to cluster web pages that have the similar contents using the metric. Fetterly et al. [6] used the metric to investigate how web pages evolve.

$$D_{k-SH}(p, p') = 1 - \frac{|S(p) \cap S(p')|}{|S(p) \cup S(p')|} \quad (4)$$

The shingling metric is oversensitive to the changes on small web pages. For example, assume that a web page <w_1, w_2, w_3, w_4, w_5> changes to <w_1, w_2, w_3, w_6, w_5>, and the shingle size k is 3. Even though only one word (w_4) changes, the metric returns 0.75. The result could cause users to misinterpret to mean that 75% of the page has changed. Moreover, if w_1 moves between w_3 and w_4 on a page <w_1, w_2, w_3, w_4, w_5>, the metric returns 1.

3 Criteria for Evaluating the Metrics

In this section, we define criteria for web page changes to evaluate how effectively the metrics measure the degree of web page changes. Prior to defining the criteria, we classify web page changes into six types (namely, "add", "copy", "drop" "shrink", "replace", and "move"). Examples of the six change types are given in Fig. 1.

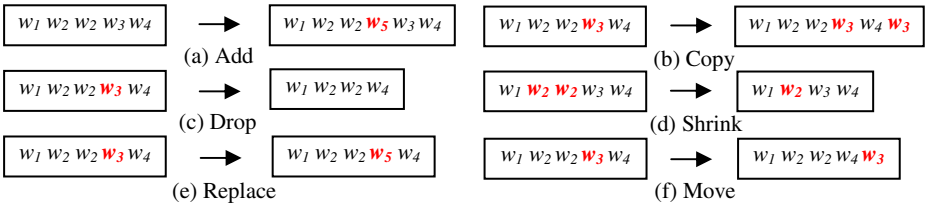


Fig. 1. Six types of web page changes

Definition 1. When new words (i.e., words not occurring on p) are inserted into p, we say that an "*add*" change takes place on the page. When old words (i.e., words occurring on p) are inserted into p, we say that a "*copy*" change takes place on the page.

Definition 2. When unique words (i.e., words occurring only once on p) are deleted from p, we say that a "*drop*" change takes place on the page. The deleted words do not exist any more on p' after the "drop" change. When duplicate words (i.e., words occurring more than once on p) are deleted from p, we say that a "*shrink*" change takes place on the page. The deleted words still occur on p' even after the "shrink" change.

Definition 3. When words on p are substituted by different words, we say that a "*replace*" change takes place on the page.

Definition 4. When the positions of words on p change, we say that a "*move*" change takes place on the page.

We give an example to illustrate how the six change types are applicable. Suppose an e-commerce site that displays a list of book information such as titles, summaries, prices, popularities (i.e., the total number of books sold), and customers' opinions, in

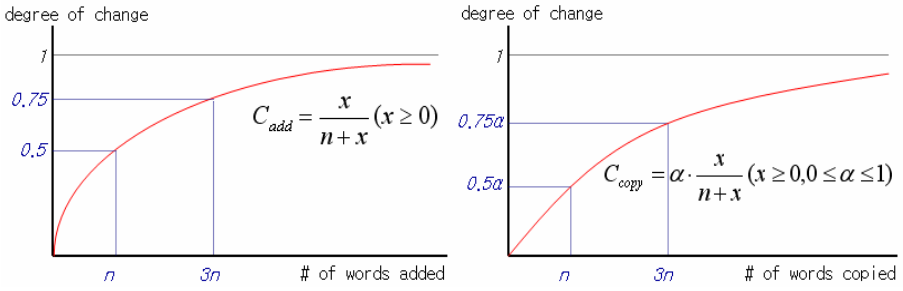


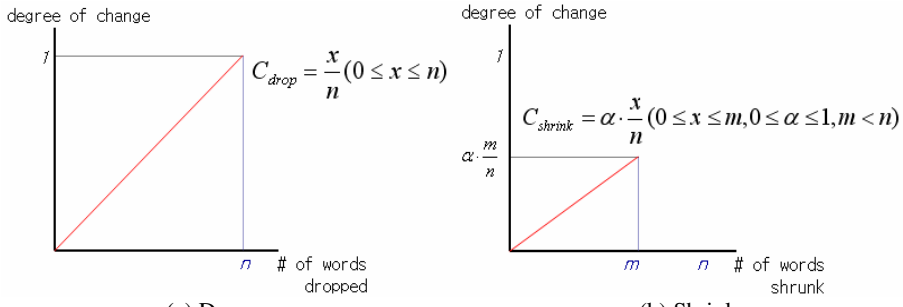
Fig. 2. Criteria for "Add" and "Copy"

an order of popularity. First, suppose that a customer writes his opinion on the page. When the opinion is different from existing opinions, the inserted words are likely to be unique (i.e., "add" change). When the opinion is similar to existing opinions, the inserted words are likely to be duplicated (i.e., "copy" change). Next, suppose that a customer deletes his opinion from the page. When the opinion is unique from existing ones, the deleted words are likely to be absent on the page (i.e., "drop" change). When the opinion is similar to other ones, the deleted words are still likely to occur on the page (i.e., "shrink" change). We distinguish the "add" change (the "drop" change) from the "copy" change (the "shrink" change, respectively), because the change of unique information is more significant than the change of duplicate information in general. On the price change, the old price of the book is updated by the new one on the page (i.e., "replace" change). When the popularities of books change, the order of book information in the list is changed (i.e., "move" change).

Figs. 2 to 4 illustrate the criterion of each change type. Let n and x denote the number of words on p and the number of changed words, respectively. The x-axis represents the number of changed words on a web page and the y-axis represents the change degree of a page.

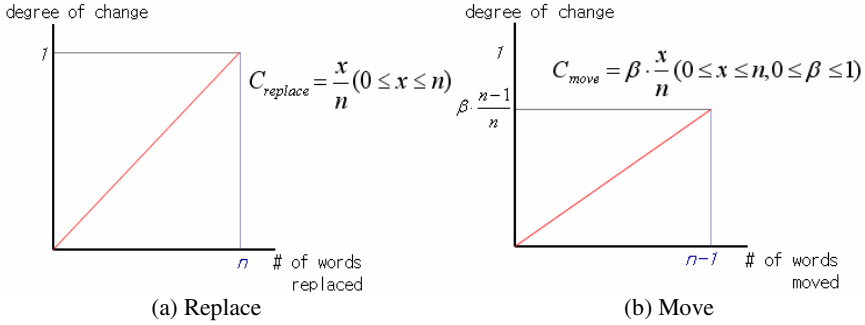
The criterion for the "add" change is defined as $(x / (n+x))$, as illustrated in Fig. 2(a). For example, when n words are added to p with n words, the change degree is 0.5 (= $n / (n+n)$). Similarly, when $3n$ words are added to p with n words, the change degree is 0.75 (= $3n / (3n+n)$). As the number of added words increases, the change degree becomes to be close to 1. The criterion for the "copy" change is defined as $(\alpha x / (n+x))$, which is illustrated in Fig. 2(b). The parameter α, which ranges from 0 to 1, denotes the user-defined weight of the "copy" change against the "add" change. As a user considers the "copy" change more significantly (or trivially), α becomes higher (or lower, respectively). For example, if a user considers the effect of adding one word to be equivalent to the effect of copying two words, α should be set to be 1/2. If a user considers the effect of adding one word to be equivalent to the effect of copying three words, α should be set to be 1/3.

The criterion for the "drop" change is defined as (x / n), as in Fig. 3(a). For example, when n words are dropped from p with n words, the degree of change is one (= n / n). More than n words cannot be dropped from a page with n words. The criterion for the "shrink" change is defined as $(\alpha x / n)$, as in Fig. 3(b). The parameter α, which is defined before, is used to denote the user-defined weight of the "shrink"



(a) Drop (b) Shrink

Fig. 3. Criteria for "Drop" and "Shrink"



(a) Replace (b) Move

Fig. 4. Criteria for "Replace" and "Move"

change against the "drop" change. m denotes the maximum number of duplicate words on p, hence more than m words cannot be shrunk.

The criterion for the "replace" change is defined as (x / n), as shown in Fig. 4(a). For example, when n words on p with n words are replaced to other words, the degree of change is one (= n / n). The criterion for the "move" change is defined as $(\beta x / n)$, as in Fig. 4(b). The parameter β, which ranges from 0 to 1, denotes the user-defined weight of the "move" change against the "replace' change. As a user considers the "move" change more significantly (or trivially), β becomes higher (or lower, respectively). $(n-1)$ is the maximum number of movable words on a page with n words.

4 A New Metric

In this section, we propose an effective metric. Our metric is an improved version of the edit distance metric. The metric considers all the six change types described in section 3. Note that the edit distance is the least expensive cost for sequences of edit operations (generally, "insertion", "deletion", and "substitution") required to transform one sequence to another sequence. In our metric, we extend the edit operations to be the following six ones: "add", "drop", "copy", "shrink", "replace", and "move". The cost of "add", "drop", and "replace" operations is 1. The cost of

"copy" and "shrink" is a, and the cost of "move" operation is b, where a and b range from 0 to 1. The extended edit distance δ_E of two web pages is defined as equation (5), where each page is regarded as a word sequence as done in [8]. k means the maximum number of sequences of the extended edit operations, and $COST_i(op)$ denotes the cost of the given edit operation in the i^{th} sequence.

$$\delta_E = \min_{i=1..k} \left\{ \sum_{op \in \{add,\ drop,\ copy,\ shrink,\ replace,\ move\}} COST_i(op) \right\} \quad (5)$$

The extended edit distance δ_E between two word sequences A and B can be obtained through the following steps (see Example 1).

Phase 1: Finding the longest common subsequence

We first find the longest common subsequence $LCS(A, B)$ of two word sequences, A and B. A subsequence of a sequence is simply a sequence with some elements (possibly none) left out. For example, a sequence $<w_2, w_3, w_4, w_2>$ is a subsequence of a sequence $<w_1, w_2, w_3, w_2, w_4, w_1, w_2>$ with corresponding index sequence $<2, 3, 5, 7>$ [5]. Using $LCS(A, B)$, we create two word sequences excluding $LCS(A, B)$ from A and B (hereafter referred to as A' and B', respectively).

Phase 2: Taking care of the "move" operations

The candidate words for "move" operation (referred to as CW_{move}) are the common words on both A' and B'. If b is bigger than $2a$, we exclude some words occurring more than once on both A and B, from CW_{move}. The excluded words become the candidate words for "copy" and "shrink" operations in the next phase so as to minimize the cost of the edit operation sequence. Then, remained CW_{move} are considered as moved words. The number of "move" operations becomes the number of moved words, and the moved words are removed from A' and B'.

Phase 3: Taking care of the "copy" and "shrink" operations

The words that exist on B' and occur on B more than once are the candidates for "copy" operation (referred to as CW_{copy}). And the words that exist on A' and occur on A more than once are the candidate words for "shrink" operation (referred to as CW_{shrink}). First, each of the common words in CW_{copy} and CW_{shrink} is considered as a shrunk and copied word; the common words exist only when b is bigger than $2a$. The numbers of "copy" and "shrink" operations become the number of the common words, and the words are removed from A', B', CW_{copy}, and CW_{shrink}. Second, note that one "copy" and one "shrink" operations may also be represented by one "replace" operation. If a is smaller than 0.5, the cost (i.e., $2a$) of one "copy" operation and one "shrink" operation is cheaper than the cost (i.e., 1) of one "replace" operation. Hence, we find words such that a word in CW_{copy} and a word in CW_{shrink} occur on the same index of B and A respectively. Each of the detected words is considered as a shrunk and copied word. The numbers of "copy" and "shrink" operations increase as many as the number of the detected words, and the words are removed from A', B', CW_{copy}, and CW_{shrink}. If a is bigger than 0.5, we exclude the detected words from CW_{copy} and CW_{shrink}. The excluded words are considered as

replaced words in the next phase. Finally, remained CW_{copy} and CW_{shrink} are considered as copied and shrunk words respectively. Hence, the numbers of "copy" and "shrink" operations increase as many as the numbers of copied words and the number of shrunk words respectively. The copied words are removed from *B'* and the shrunk words are removed from *A'*.

Phase 4: Taking care of the "replace", "add" and "drop" operations

The words on *A'* and *B'* with the same indexes on both *A* and *B* are considered as replaced words. The number of "replace" operations becomes the number of replaced words and the replaced words are removed from *A'* and *B'*. Next, the words on *A'* and *B'* are considered as dropped and added words respectively. Hence the numbers of "add" and "drop" operations become the numbers of the words on *B'* and *A'* respectively.

Phase 5: Calculating the extended edit distance

The extended edit distance from *A* to *B* is calculated from the numbers of edit operations and their costs.

Example 1. Suppose that a web page $A = <w_1, w_2, w_2, w_2, w_3, w_3, w_4, w_5, w_2>$ changes to $B = <w_3, w_1, w_4, w_2, w_3, w_5, w_5, w_6, w_6, w_7>$. The parameters *a* and *b* are set to be 0.4 and 0.9, respectively. We compute the extended edit distance from *A* to *B* as follows.

At the first phase, we find the longest common subsequence of *A* and *B*:

$LCS(A, B) = <w_1, w_2, w_3, w_5>$
$A' = <w_2, w_2, w_3, \mathbf{w_4}, w_2>$
$B' = <w_3, \mathbf{w_4}, w_5, w_6, w_6, w_7>$
|add| = 0, |drop| = 0, |copy| = 0, |shrink| = 0, |replace| = 0, |move| = 0

At the second phase, CW_{move} are w_3 and w_4 since the two words occur on both *A'* and *B'*. But, since *b* is bigger than *2a* and w_3 occurs twice on both *A* and *B*, w_3 is excluded from CW_{move}. Hence, only w_4 is considered as a moved word. The number of "move" operations becomes one and w_4 is removed from *A'* and *B'*:

$A' = <w_2, w_2, \mathbf{w_3}, w_2>$
$B' = <\mathbf{w_3}, w_5, w_6, w_6, w_7>$
|add| = 0, |drop| = 0, |copy| = 0, |shrink| = 0, |replace| = 0, |move| = 1

At the third phase, CW_{copy} are w_3, w_5, and w_6, and CW_{shrink} are w_2, w_2, w_2, and w_3. First, w_3 is considered as a shrunk and copied word because the word is a common word in CW_{copy} and CW_{shrink}. Hence, the number of "copy" and "shrink" operations becomes one and w_3 is removed from *A'*, *B'*, CW_{copy}, and CW_{shrink}. At this time, we have:

$A' = <w_2, w_2, \mathbf{w_2}>$, $CW_{shrink} = w_2, w_2, w_2$
$B' = <w_5, w_6, \mathbf{w_6}, w_7>$, $CW_{copy} = w_5, w_6$
|add| = 0, |drop| = 0, |copy| = 1, |shrink| = 1, |replace| = 0, |move| = 1

Next, we find the words with the same indexes on both *A* and *B* from CW_{copy} and CW_{shrink}, which are w_2 and w_6; w_2 occurs on *A* with the ninth index and w_6 occurs on *B* with the ninth index. Since *a* is smaller than 0.5, w_2 and w_6 are considered as a shrunk word and a copied word respectively. Hence, the numbers of "copy" and "shrink"

operations increase by one. w_2 is removed from A' and CW_{shrink}, and w_6 is removed from B' and CW_{copy}:

$A' = \langle \boldsymbol{w_2}, \boldsymbol{w_2} \rangle$, $CW_{shrink} = w_2, w_2$
$B' = \langle \boldsymbol{w_5}, w_6, w_7 \rangle$, $CW_{copy} = w_5$
|add| = 0, |drop| = 0, |copy| = 2, |shrink| = 2, |replace| = 0, |move| = 1

Next, remained CW_{copy} and CW_{shrink} are considered as copied words and shrunk words respectively. Hence, the number of "copy" operations increases by one and the number of "shrink" operations increases by two. w_2 and w_2 are removed from A', and w_5 is removed from B':

$A' = \langle \rangle$
$B' = \langle w_6, w_7 \rangle$
|add| = 0, |drop| = 0, |copy| = 3, |shrink| = 4, |replace| = 0, |move| = 1

At the fourth phase, there are no any replaced or dropped words since A' is empty. w_6 and w_7 are considered as added words, hence the number of "add" operations becomes two. Finally we have the numbers of all the extended edit operations in the least expensive sequence of the operations:

|add| = 2, |drop| = 0, |copy| = 3, |shrink| = 4, |replace| = 0, |move| = 1

At the fifth phase, we calculate the extended edit distance from A to B:

$\delta_E = (2 \cdot 1) + (0 \cdot 1) + (3 \cdot 0.4) + (4 \cdot 0.4) + (1 \cdot 0) + (1 \cdot 0.9) = 5.7$ ■

Using the extended edit distance, we calculate the distance between two web pages p and p' as equation (6), where m and n denote the numbers of words on p and p' respectively.

$$D_{IED}(p, p') = \frac{\delta_E}{\max(m, n)} \tag{6}$$

5 Experiments

We conducted two experiments. First, using real web pages, we show how differently each similarity metric measures the degree of web page changes. Second, we present the effectiveness of the metrics under the proposed criteria. We compare the following six metrics: the byte-wise comparison (in short, BW), the TF·IDF cosine distance (COS), the word distance (WD), the edit distance (ED), the 10-shingling (10SH), and our metric improving the edit distance metric (IED). Markups of web pages were excluded in the experiments, as done in the literature [2, 6, 8, 9]. The parameters a and b of our metric were set to be 0.75.

5.1 Difference of the Metrics

We randomly crawled 41,469 Korean web pages in August 2005. The web pages were downloaded twice in a two-day interval. From two versions of the web pages, we measured the degree of changes using the six metrics. Fig. 5 shows the change

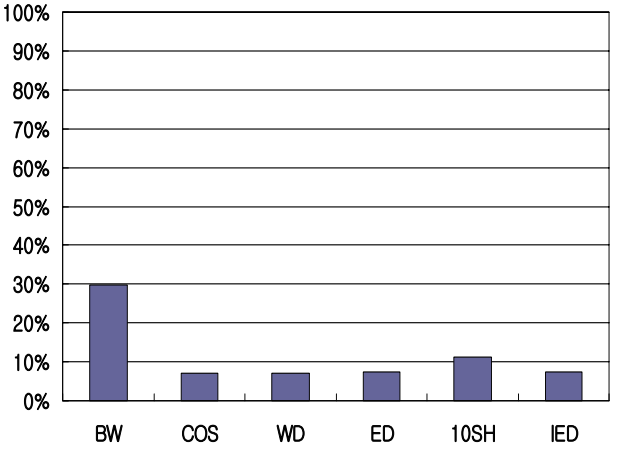
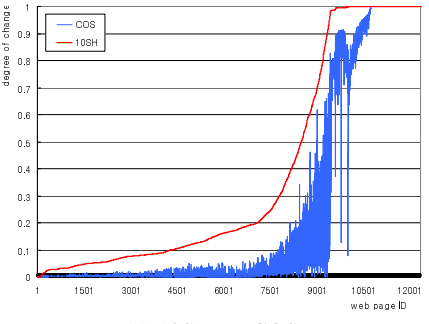


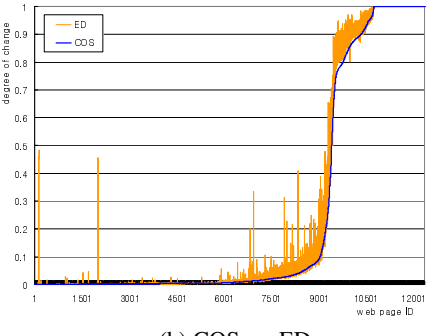
Fig. 5. Difference of sensitivity (I)

degrees of the pages under the six metrics. *y*-axis represents the sum of the change degrees of all the pages as percentage. The BW result implies that 30% of the pages changed, since it always returns 1 despite of very trivial changes (say, insertion of a blank space). For the same 30% of the pages, other metrics respond differently. 10SH determines that about 12% of all the page contents changed, while COS, WD, ED, and IED say that about 7% of them changed. As Fig. 5 shows, BW is the most sensitive metric among the six metrics. We also learn that 10SH is more sensitive to the page changes than COS, WD, ED, and IED are.

Fig. 6 shows how differently 10SH, COS, and ED respond to the same changes of web pages. The *x*-axis represents the identifier of each web page, and the *y*-axis represents the change degree of the corresponding web page. The identifiers of web pages are sorted in an ascending order of 10SH and COS in Fig. 6(a) and 6(b) respectively, in order to clearly visualize the difference of the metrics. We have observed the similar results for other combinations of metrics, which are not explicitly



(a) 10SH vs. COS



(b) COS vs. ED

Fig. 6. Difference of sensitivity (II)

presented though. As the figures show, each metric returns different values on almost all the pages. Sometimes, the difference is as large as 0.92 as shown in Fig. 6(a). This experiment implies that the degree of web page changes is heavily dependent on which metric is used to measure the page changes. Users need to select an appropriate metric when they precisely measure the degree of web page changes; otherwise they would misunderstand web page changes. These experimental results motivated our study.

5.2 Evaluation of the Metrics

We conducted experiments to evaluate the effectiveness of the metrics under the proposed criteria. These experiments were done with synthesized pages, which are constructed to reflect the characteristics on the web. A metric is defined to be effective if the results of the metric are close to the criteria. In addition, if the results of a metric are always higher (or lower) than the criteria, we say that the metric is oversensitive (or undersensitive, respectively). α and β in the criteria were set to be 0.75.

First, we evaluated the metrics when various numbers of words on a page with 1,000 words were changing. We chose a page with 1000 words, since web pages with 1,000 words occupy about 25% on the web [6]. The changed words were clustered (i.e., not distributed) on the pages, because the changes of real web pages were generally clustered [8]. In Fig. 7, the *x*-axis represents the number of changed words on a page, and the *y*-axis denotes the change degree of the corresponding page. 10SH is effective for the "add" and "drop" changes, but is oversensitive for the other changes. If α in the criteria were set to be one, the metric would be effective for the "copy" and "shrink" changes. In our experiment, COS is always undersensitive. COS returns very low values for the "copy" and "shrink" changes, which implies that COS treats the "copy" and "shrink" changes to be minor. COS and WD always return zero on the "move" change because they do not consider the changes of word order at all. WD is effective for the "replace" change but is undersensitive for the other changes. If α in the criteria were to be 0.5, the metric would be effective for the "copy" change. On the other words, WD would be the right choice for users who consider the effect of adding one word to be equivalent to the effect of copying two words. ED works similar to WD, except for the "move" change. ED treats the "move" change and "replace" change identically. IED returns the most effective results in all cases.

Next, we evaluated the metrics on various sizes of pages (i.e., 2^2, 2^3, 2^4, … , or 2^{13} words). Note that web pages with 2^2 to 2^{13} words occupy about 95% on the web [6]. We maintained the fraction of changed words on each page to be 1/4; one word on a 2^2 word page, two words on a 2^3 word page, three words on a 2^4 word page, and so on. The *x*-axis in Fig. 8 represents the number of words on a page before change. From this result, we found out that 10SH becomes more oversensitive as web pages become smaller on all the change types. The sensitivities of the other metrics are not dependent to the page size in most cases; COS also varies in sensitivity according to the page size, but it is not serious.

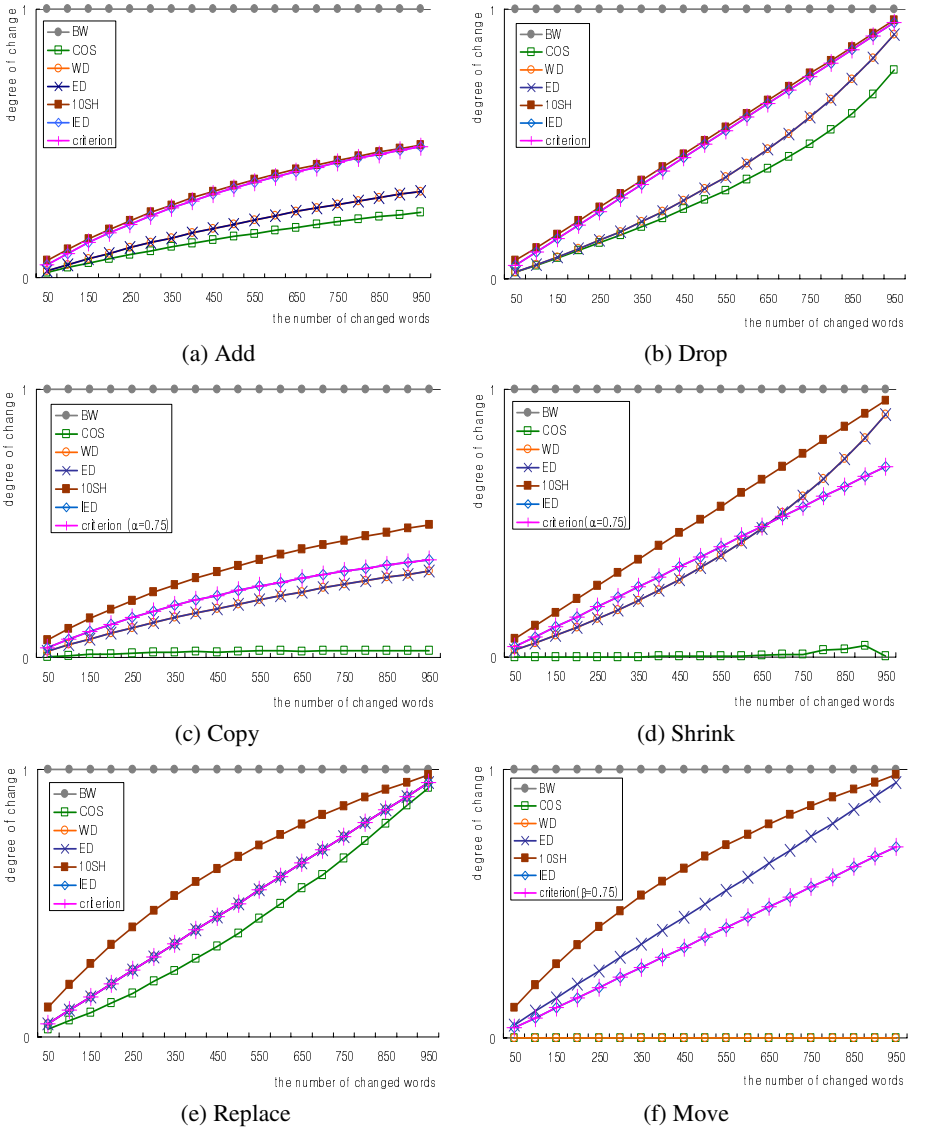
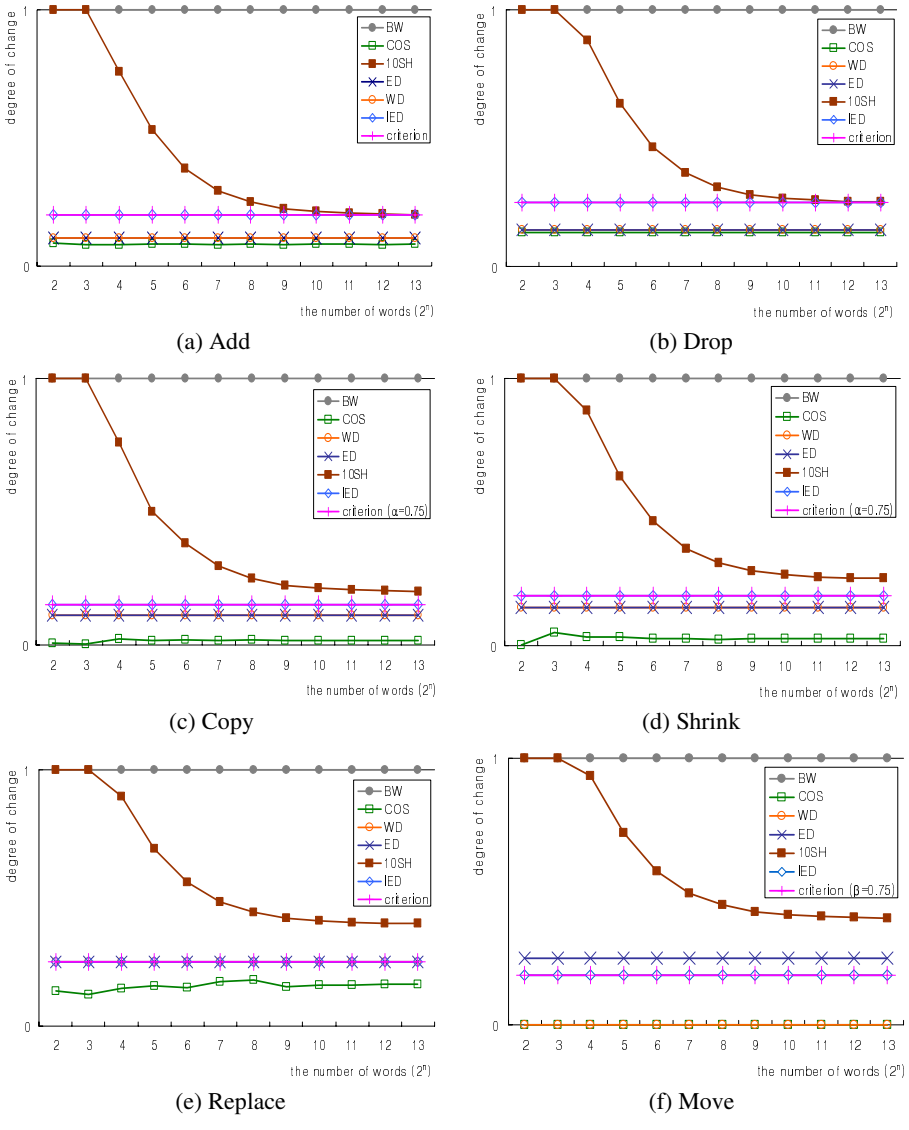


Fig. 7. Comparison of metrics

Note that the experiments were done with $\alpha = 0.75$ and $\beta = 0.75$. If parameters α and β were set differently, the sensitivities (or effectiveness) of the metrics would be differently evaluated under the "copy", "shrink", and "move" changes. It should be observed that the criteria graphs with other values of α and β could be easily predicted; only the slope of the criteria graph changes.



(a) Add (b) Drop

(c) Copy (d) Shrink

(e) Replace (f) Move

Fig. 8. Sensitivity versus page size

We now summarize what we have learned in our experiments.

1. 10SH is oversensitive to the changes of web pages (especially under the "move" change). The smaller the page size is, the more sensitive 10SH is. 10SH is effective for the "add" and the "drop" changes. 10SH measures the "add" change and the "copy" change similarly. The "drop" and "shrink" changes are similarly treated under 10SH.

2. COS is an undersensitive metric. COS does not consider the "move" change at all. In particular, the "copy" and the "shrink" changes are regarded as minor changes.
3. WD does not consider the "move" change at all. WD is undersensitive to the "add" and the "drop" changes. The "copy" changes (or the "shrink" changes) of two words are approximately regarded as the "add" change (or the "drop"change, respectively) of one word. WD is well effective for the "replace"change.
4. ED shows the same results as WD except the "move" change. The "move" change is considered similarly with the "drop" or "replace" change.
5. IED is the most effective metric under our proposed criteria. Mostly, IED responds to the changes of web pages identically to the criteria. IED can be used effectively in various applications because it can adjust the weights of the "copy", "shrink", and "move" changes.

6 Closing Remarks

In this paper, we classified the changes of web pages into six types, which are "add", "copy", "drop" "shrink", "replace", and "move", then we defined a criterion for each type. We also proposed a new metric designed to reflect the criteria well. Under the criteria, we evaluated the effectiveness of the six metrics (namely, the byte-wise comparison, the TF·IDF cosine distance, the word distance, the edit distance, the shingling, and ours). Based on this evaluation, we found that the proposed metric is the most effective metric in terms of all the six change types. Our study presents how significantly the metrics consider each change type and which metric is effective on each change type. We believe that this study is the first attempt to evaluate the metrics and could be used as a guideline for selecting an appropriate metric measuring the degree of web page changes.

With our criteria, the metrics can be evaluated separately in terms of one of the six change types. Even though more than one change type often occurs simultaneously on real web pages, our criteria have some limitations to measure multiple changes. Further research on the criteria for evaluating which metrics represent simultaneous changes effectively is necessary.

References

1. Brewington, B. E., Cybenko, G.: How Dynamic is the Web? the 9th International World Wide Web Conference (2000) 257-276
2. Broder, A. Z., Glassman, S. C., Manasse, M. S., Zweig, G.: Syntactic Clustering of the Web. Computer Networks and ISDN Systems Vol. 29. No. 8-13. (1997) 1157-1166
3. Cho, J., Garcia-Molina, H.: The Evolution of the Web and Implications for an Incremental Crawler. the 26th International Conference on Very Large Data Bases (2000) 200-209
4. Cho, J., Garcia-Molina, H.: Synchronizing a Database to Improve Freshness. the ACM SIGMOD International Conference on Management of Data (2000) 117-128
5. Cormen, T. H., Leiserson, C. E., Rivest, R. L.: Introduction to Algorithm. the Massachusetts Institute of Technology (2001)

6. Fetterly, D., Manasse, M., Najork, M., Wiener, J. L.: A Large-Scale Study of the Evolution of Web Pages. Software: Practice & Experience, Vol. 34, No. 2 (2003) 213-237
7. Kim, S. J., Lee, S. H.: An Empirical Study on the Change of Web Pages. the 7th Asia Pacific Web Conference (2005) 632-642
8. Lim, L., Wang, M., Padmanabhan, S., Vitter, J. S., Agarwal, R.: Characterizing Web Document Change. the 2nd International Conference on Advances in Web-Age Information Management (2001) 133-144
9. Ntoulas, A., Cho, J., Olston, C.: What's New on the Web? The Evolution of the Web from a Search Engine Perspective. the 13th International World Wide Web Conference (2004) 1-12
10. Salton, G., McGill, M. J.: Introduction to Modern Information Retrieval. McGraw-Hill (1983)