

Information Shared by Many Objects

Chong Long*, Xiaoyan Zhu†, Ming Li‡, Bin Ma§

ABSTRACT

If Kolmogorov complexity [25] measures information in one object and Information Distance [4, 23, 24, 42] measures information shared by two objects, how do we measure information shared by many objects? This paper provides an initial pragmatic study of this fundamental data mining question. Firstly, $E_m(x_1, x_2, \dots, x_n)$ is defined to be the minimum amount of thermodynamic energy needed to convert from any x_i to any x_j . With this definition several theoretical problems have been solved. Second, our newly proposed theory is applied to select a comprehensive review and a specialized review from many reviews: (1) Core feature words, expanded words and dependent words are extracted respectively. (2) Comprehensive and specialized reviews are selected according to the information among them. This method of selecting a single review can be extended to select multiple reviews as well. Finally, experiments show that this comprehensive and specialized review mining method based on our new theory can do the job efficiently.

Keywords

data mining, text mining

1. INTRODUCTION

A great deal of data mining research can be regarded as gathering information from information carrying objects. However, without a general agreement of what is information in one object, what is information shared by two objects, and what is information shared among many objects, we end up

*Department of Computer Science, Tsinghua University, Beijing, China. longc05@mails.tsinghua.edu.cn

†Department of Computer Science, Tsinghua University, Beijing, China. zxy-dcs@tsinghua.edu.cn

‡Corresponding author. David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada, mli@uwaterloo.ca

§Corresponding author. Department of Computer Science, University of Western Ontario, London, Ontario, Canada, bma@uwo.ca

with dozens of arbitrary measures and algorithms that perhaps optimal under one measure, but not under another. The authors of [39] have articulated this problem. The field will certainly continue to grow and flourish without settling such a problem, as it has been, as a field of engineering.

This work represents another step of our continued efforts to solve this problem. Over the past decade, we have answered the question of “what is the shared information between two objects”. We introduced the metric of information distance [4, 23, 24] that is provably better than all other “reasonable” metrics in all application domains. These include all metrics listed in [39] that satisfy distance metric requirements and when they are normalized to the range of 0 and 1 to be compared with the normalized information distance. This theory has been widely accepted and further studied by the theoretical community [41, 40, 36, 27, 26, 9]. Our theory has also led to several successful applications in the data mining community, for example, [18, 42]. In [18], Keogh, Lonardi, and Ratanamahatana compared a variant of our approach in [23] to 51 measures from 7 data mining related conferences including SIGKDD, SIGMOD, ICDM, ICDE, SSDB, VLDB, PKDD, PAKDD, and have concluded that our information distance based method was superior to all these parameter-laden methods on their benchmark data. In the meantime, the theory has found dozens of applications in many fields from weather forecasting to software engineering, and to bioinformatics [1, 3, 8, 11, 12, 10, 14, 19, 21, 20, 22, 38, 30, 31, 32, 35, 2, 34, 28, 29]. A complete list of references is in the third edition of [25].

However, in many data mining applications, we are more interested in mining shared information from many, not just two, information carrying entities. For example, what is the public opinion on the United States presidential election, from the blogs? What do the customers say about a product, from the reviews? Which article, among many, covers the news most comprehensively? Or specialized in one particular news item?

Kolmogorov complexity and our prior theory of information distance are not sufficient for these tasks. Kolmogorov complexity deals with one object and information distance deals with two objects. There is a conspicuous gap: a theory dealing with many objects.

2. PRELIMINARIES

2.1 Information in one object

Kolmogorov complexity was introduced almost half a century ago by R. Solomonoff, A.N. Kolmogorov and G. Chaitin, see [25]. It is now widely accepted as an information theory for individual objects parallel to that of Shannon’s information theory which is defined on an ensemble of objects. Fix a universal Turing machine U . The Kolmogorov complexity of a binary string x condition to another binary string y , $K_U(x|y)$, is the length of the shortest (prefix-free) program for U that outputs x with input y . It can be shown that for different universal Turing machine U' , for all x, y

$$K_U(x|y) = K_{U'}(x|y) + C,$$

where the constant C depends only on U' . Thus we simply write $K_U(x|y)$ as $K(x|y)$. We write $K(x|\epsilon)$, where ϵ is the empty string, as $K(x)$. For a comprehensive study of Kolmogorov complexity and its applications, see [25].

2.2 Information Distance between two objects

In the classical Newton’s world, “distance” is measured uniquely. This has not been the case for distance in cyber space. A good information distance metric should not only be application independent but also provably better than other “reasonable” definitions.

Traditional distances such as the Euclidean distance or the Hamming distance fail for even trivial examples. Tan et al [39] have demonstrated that none of the 21 metrics used in data mining community is universal, practically. In fact, for any computable distance, we can always find counterexamples.

What would be a good departure point for defining an “information distance” between two objects? To answer this question, in the early 1990’s, we [4] have studied the energy cost of conversion between two strings x and y . John von Neumann hypothesized that performing 1 bit of information processing costs $1KT$ of energy, where K is the Boltzmann’s constant and T is the room temperature. Observing that reversible computations can be done for free, in early 1960’s Rolf Landauer revised von Neumann’s proposal to hold only for irreversible computations. We proposed in [4] to use the minimum energy needed to convert between x and y to define their distance, as it is an objective measure. Thus, if one wishes to erase string x , then one can reversibly convert it to x^* , x ’s shortest effective description, then erase x^* . Only the process of erasing $|x^*|$ bits is irreversible computation. Carrying on from this line of thinking, we have defined in [4] that the energy to convert between x and y to be the smallest number of bits needed to convert from x to y and vice versa. That is, with respect to a universal Turing machine U , the cost of conversion between x and y is:

$$E(x, y) = \min\{|p| : U(x, p) = y, U(y, p) = x\} \quad (1)$$

It is clear that $E(x, y) \leq K(x|y) + K(y|x)$. From this observation, and some other concerns, we have defined the sum distance in [4]:

$$D_{\text{sum}}(x, y) = K(x|y) + K(y|x).$$

However, the following theorem proved in [4] was a surprise.

THEOREM 1. $E(x, y) = \max\{K(x|y), K(y|x)\}$.

Thus, the max distance was defined in [4]:

$$D_{\text{max}}(x, y) = \max\{K(x|y), K(y|x)\}.$$

Both distances are shown to satisfy the basic distance requirements such as positivity, symmetricity, triangle inequality, in [4]. It was further shown that D_{max} and D_{sum} minorize (up to constant factors) all other distances that are computable and satisfies some reasonable density condition that within distance k to any string x , there are at most 2^k strings. Formally, a distance D is admissible if

$$\sum_y 2^{-D(x,y)} \leq 1. \quad (2)$$

$D_{\text{max}}(x, y)$ satisfies the above requirement because of Kraft’s Inequality (with the prefix-free version of Kolmogorov complexity). It was proved in [4] that for any admissible computable distance D , there is a constant c , for all x, y ,

$$D_{\text{max}}(x, y) \leq D(x, y) + c. \quad (3)$$

Putting it bluntly, if any such distance D discovers some similarity between x and y , so will D_{max} .

Then, after normalization [23] and [24],

$$d_{\text{max}}(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}, \quad (4)$$

this theory has been initially applied to alignment free whole genome phylogeny [23], chain letter history [5], language history [3, 24], plagiarism detection [8], and more recently to music classification and clustering [11, 10], parameter-free data mining paradigm [18], protein sequence classification [20], protein structure comparison [16], heart rhythm data analysis [37, 35], question and answering system [42], bioinformatics [28, 29], and many more. However, in many of these applications, even when many objects are involved, only pair-wise information distance is computed. This has limited the applicability of this theory.

3. SHARED INFORMATION AMONG MANY OBJECTS

Can we generalize the theory of information distance to more than two objects? We make a new proposal in this section and perform experiments in the next section.

Similar to Formula 1, given strings x_1, \dots, x_n , we can define the minimum amount of thermodynamic energy needed to convert from any x_i to any x_j as:

$$E_m(x_1, \dots, x_n) = \min\{|p| : U(x_i, p, j) = x_j \text{ for all } i, j\} \quad (5)$$

Clearly,

$$E_m(x_1, \dots, x_n) \leq \sum_i K(x_1 x_2 \dots x_n | x_i).$$

However, similar to Theorem 1, the following theorem demonstrates a rather surprising property.

THEOREM 2. *Modulo to an $O(\log n)$ additive factor,*

$$E_m(x_1, \dots, x_n) = \max_i K(x_1 x_2 \dots x_n | x_i)$$

PROOF. Suppose all binary strings are given in a list s_1, s_2, \dots . Define a set V as follows: a vector $v = (i_1, i_2, \dots, i_n)$ is in V if and only if $K(s_{i_1} s_{i_2} \dots s_{i_n} | s_{i_j}) \leq K(x_1 x_2 \dots x_n | x_j)$ for every $j = 1, 2, \dots, n$.

Regard V as the vertices of a graph $G = \langle V, E \rangle$. Two vertices $u = (u_1, \dots, u_n)$ and $v = (v_1, \dots, v_n)$ are such that $(u, v) \in E$ if and only if there is $1 \leq j \leq n$ such that $u_j = v_j$. For any given $u \in V$ and $1 \leq j \leq n$, by the definition of V , there are at most $2^{K(x_1 x_2 \dots x_n | x_j)}$ vertices $v \in V$ such $v_j = u_j$. Denote $D = \max_i K(x_1 x_2 \dots x_n | x_i)$. The degree of the graph G is therefore bounded by

$$\sum_j 2^{K(x_1 \dots x_n | x_j)} \leq n \times 2^D.$$

It is known that a graph with degree d has a d -coloring. Therefore, G has a coloring $V = V_1 \cup V_2 \cup \dots \cup V_K$ such that $K \leq n \times 2^D$. Clearly, $(x_1, x_2, \dots, x_n) \in V$. In order to compute x_i from x_j for any pair of i and j , a universal turing machine only needs to know which V_k contains (x_1, x_2, \dots, x_n) . Such a program needs only $\log_2(n \times 2^D) = D + \log_2 n$ bits. \square

Notice that Theorem 2 is a strong claim. Comparing to Theorem 1 where the saving is only linear, the saving here is quadratic. It is possible to prove that E_m satisfies the usual metricity properties such as being symmetric and satisfying the triangle inequality. It is also possible to prove the E_m also has the ‘‘universality’’ property that for any other non-trivial distance E'_m for many objects satisfying the above properties, we always have for all x_1, \dots, x_n , we have $E_m(x_1, \dots, x_n) \leq E'_m(x_1, \dots, x_n) + O(1)$. Thus this provides a theory guiding us to compute how much information a given set of n objects share.

The following theorem is a corollary of Theorem 2:

THEOREM 3. *Modulo to an $O(\log n)$ additive factor,*

$$\min_i K(x_1 \dots x_n | x_i) \leq E_m(x_1, \dots, x_n) \leq \min_i \sum_{k \neq i} D_{\max}(x_i, x_k). \quad (6)$$

Given n objects, the left-hand side of the equation may be interpreted as the most comprehensive object that contains the most information about all of the others. The right-hand side of the equation may be interpreted as the most specialized object that is similar to all of the others.

Let us consider news items on the internet. If we wish to choose a news article that covers the most news, we can use the left hand side of Theorem 3. If we wish to look at a typical coverage of a single topic, we can use the right-hand side of Theorem 3.

In the next section, we will use this new theory to guide our practical work. The easiest-to-obtain dataset for experimenting our theory turns out to be product reviews. These data can be easily annotated, too. As it turns out, our work also provides an interesting fresh perspective to the work of summarization.

4. REVIEW SELECTION USING INFORMATION DISTANCE

4.1 Comprehensive and Specialized Reviews

With the rapid development of Web2.0 and e-commerce that emphasizes the participation of users, more and more Websites, such as Amazon (<http://www.amazon.com>) and Epinions (<http://www.epinions.com>), encourage people to express opinions on products by posting reviews [43]. These reviews are very useful for readers and will possibly influence their purchasing decisions. However, it would cost too much time for readers to read all of the hundreds of reviews of the same product. Thus, automatic review mining and summarization is a very practical concern. The most research on this topic focus on feature-opinion pairs extraction and sentiment orientation decision [17, 33, 15, 7].

However, a human reader is usually not completely satisfied by a machine generated dull report and may still prefer to read a vivid and complete review article written by a good human writer. This raises the need of selecting the best review from a set of reviews. If only one review from a set was to be read, the most sensible choice would be the most *comprehensive* review that covers the most information about the other reviews. This is the first goal of our review selection method.

After the most comprehensive review is read, a user may become more interested in one particular feature of the product and would like to read another concise and representative review focusing on that feature only. Therefore it becomes useful to select the best *specialized* review that focuses on a given feature and represents the other reviewers’ opinions on that feature only. This becomes the second goal of our review selection method.

Our review selection method is based on the information distance discussed in the previous section. From the discussion after Theorem 3, the left-hand side and right-hand side of Equation (6) defines a way to select the most comprehensive and the best specialized reviews, respectively. However, our problem is that neither the Kolmogorov complexity $K(\cdot, \cdot)$ nor $D_{\max}(\cdot, \cdot)$ is computable. Therefore, we have to find a way to ‘‘approximate’’ these two measures.

The most useful information in a review article is the English words that are related to the product features. If we can extract all of these related words from the review articles, the size of the word set can be regarded as a very rough estimation of information content (or Kolmogorov complexity) of the review articles. Although this is a very inaccurate approximation, in Section 5 we will see that this already gives very good practical results.

Our method is outlined in the following. First, for each type of product (such as digital camera), a small set of *core feature words* (such as price, image) is generated through statistics. Then this core set of words are used to generate the *expanded words* with an algorithm. Thirdly, an English parser is used to find the *dependent words* associated to the occurrences of the core feature words and expanded words in a review. For each review-feature pair, the union of the core feature words and expanded words in the review, and their

dependent words found by the parser define the *related word set* of the review on the feature. Lastly, each distinct word in a word set is assigned with one unit of information content; and the left-hand side and right-hand side of Formula (6) are used to select the comprehensive review and specialized review, respectively.

4.2 Word Extraction

4.2.1 Features and Core Feature Words

Here “features” broadly mean product features (or attributes) and functions that have been commented on in reviews [17]. For example, pixel, memory, shutter, battery, . . . , are features of a digital camera. Given a feature, the core feature words are the very few most common English words that are used to refer to that feature. For example, both “image” and “picture” are used to refer to the same feature of a digital camera.

Feature words are the most direct and frequent words describing a feature, therefore, if there is a feature word in a sentence, it is most likely talking about this feature.

In [17], the authors indicated that when customers comment on product features, the words they use converge. According to the statistical results of our training corpus, we can get the same conclusion. If we remove the feature words with frequency lower than 1% of the total frequency of all feature words, the remaining words, which are just core feature words, can still cover more than 90% occurrences. Then some of those with the same meaning (such as “image” and “picture”) are grouped into one feature. In our experiments, on average, each product’s each feature has 1.4 core feature words.

4.2.2 Expanded Words

Apart from core feature words, many others less-frequently used words that are connected to the feature also contribute to the information content of the feature. For example, “price” is an important feature of a product, but the word “price” is usually dropped from a sentence. Instead, words such as “\$”, “dollars”, “USD”, and “CAD” are used. It would be impossible to manually enumerate all these different expressions of the same thing. Therefore, these expanded words should be generated automatically.

We use information distance to expand words. In [12], the Google code of length $G(x)$ represents the shortest expected prefix-code word length of the associated Google event x . Then the Google distribution can be used as a compressor for the Google semantics associated with the search terms. Normalized Google distance (NGD) is defined as follows:

$$\begin{aligned} NGD(x, y) &= \frac{G(x, y) - \min(G(x), G(y))}{\max(G(x), G(y))} \\ &= \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}} \end{aligned}$$

where $f(x)$ denotes the number of pages containing x , and $f(x, y)$ denotes the number of pages containing both x and y , as reported by Google.

In our work, the distance d between words is defined according to NGD based on the words’ frequencies and co-occurrence frequencies in the training corpus. Let α be a

feature and \mathcal{A} be the set of its core feature words. The distance between a word w and the feature α is then defined to be

$$d(w, \alpha) = \min_{v \in \mathcal{A}} d(w, v).$$

Then a distance threshold is used to determine which words should be included in the set of expanded words for a given feature.

4.2.3 Dependent Words

If a core feature word or an expanded word is found in a sentence, the words which have grammatical dependent relationship [13] with it are called the dependent words. For example, in sentence “4x digital zoom is great”, the words “4x”, “digital” and “great” are all dependent words of the core feature word “zoom”. All these words also contribute to the reviews and are important to determine the reviewer’s attitude towards a feature.

The Stanford Parser [13] is used to parse each review. For review i and feature j , the core feature words and expanded words that occur in the review are first computed. Then the parsing result is examined to find all the dependent words for the core feature words and expanded words.

The word set S_{ij} is defined to be the union of all the core feature words and expanded words that occur in the review, plus all of their expanded words. Thus, a review is represented by a vector of word sets for all the product’s features: $S_i = (S_{i1}, S_{i2}, \dots, S_{in})$. Notice that S_{ij} and $S_{ij'}$ for two different features j and j' may share some common words, but with perhaps totally different meanings. For this reason, it is very important to use an English parser and keep the related word sets for different features separately.

4.3 Computing Information Distance

Let S and T be two sets of words, and each word carries one unit of information. Then the Kolmogorov complexity can be intuitively estimated by

$$K(S) = |S|, \quad K(T) = |T|, \quad \text{and} \quad K(S|T) = |S \setminus T|.$$

Here $|X|$ is the size of the set X . Such intuition can be extended to vectors of sets. For two vectors of sets $S_i = (S_{i1}, S_{i2}, \dots, S_{in})$, $i = 1, 2$, define

$$S_1 S_2 = S_1 \cup S_2 = (S_{11} \cup S_{21}, \dots, S_{1n} \cup S_{2n}),$$

$$K(S_i) = \sum_{j=1}^n K(S_{ij}),$$

and

$$K(S_1|S_2) = \sum_{j=1}^n K(S_{1j}|S_{2j}).$$

Then

$$D_{max}(S_1, S_2) = \max(K(S_1|S_2), K(S_2|S_1))$$

and $K(S_1 S_2 \dots S_n | S_i)$ can all be naturally defined as before. Thus, we are able to use Equation (6) for our review selection.

If there are m reviews x_1, x_2, \dots, x_m , and n features u_1, u_2, \dots, u_n , straightforwardly from the left-hand side of Equation (6), the most comprehensive review i is such that

$$i = \arg \min_i D_{max}(S_i, S_1 \dots S_n),$$

that is,

$$i = \arg \min_i K(S_1 \dots S_n | S_i). \quad (7)$$

The best specialized review needs some minor changes to the right-hand side of Equation (6). Without modification, the best specialized review i for a feature j would be such that

$$i = \arg \min_i \sum_k D_{max}(S_{ij}, S_{kj}).$$

However, for specialized review we want that (a) the review focuses on the given feature only, and (b) an review article that does not discuss the feature should not be counted in the selection. Therefore, the above formula is modified to be

$$i = \arg \min_i \sum_{S_{kj} \neq \emptyset} D_{max}(S_i, S_{kj}), \quad (8)$$

here

$$D_{max}(S_i, S_{kj}) = D_{max}(S_i, (\emptyset, \emptyset, \dots, S_{kj}, \dots, \emptyset))$$

where S_{kj} is in j th entry.

More specifically, S_{ij} is changed to S_i to penalize the content of review i not related to feature j ; and the reviews with an empty word set on feature j are excluded from the selection.

Our method of selecting a single review can be extended to select multiple reviews as well. For example, one can adapt the maximal marginal relevance method introduced in [6] by using our distance, and select these reviews one by one incrementally.

5. EXPERIMENTAL RESULTS

We conducted our experiments using customer reviews on two electronics products: digital cameras (DC) and televisions (TV). Totally 138,985 reviews containing 28 million words are used as the corpus for extracting features and computing information distances between feature words and other words. With these distances, expanded words are extracted by method introduced in 4.2.2.

Our experiments focus on single comprehensive and specialized review selection.

5.1 Comprehensive Review Selection

To test the performance of comprehensive review selection, six popular DC models from DC and eight TV models with most reviews are selected, resulting 14 test sets with 1381 reviews, approximately 100 reviews each ¹.

¹All these reviews, together with their annotations as well as our experimental data are available at our website <http://learn.tsinghua.edu.cn:8080/2005310464/ComprehensiveAndSpecialized.htm>.

Table 1: Top 1 Comprehensive Reviews

Product Name	No. of Sets	A	B	C	D
DC	6	3	2	1	0
TV	8	3	3	2	0
Total	14	6	5	3	0
Labeled		17	66	264	1034

Two independent teams from Canada and China, each with two people, annotated the reviews. Each review is annotated by each team, independently, with one of the following four labels according to its comprehensiveness:

1. Label ‘A’: the most comprehensive reviews. Each test set should have only one review ranked ‘A’. In a few cases, we allow a few more ‘A’s since they are equally good reviews.
2. Label ‘B’: the reviews are very close to the reviews ranked ‘A’.
3. Label ‘C’: the reviews are fairly comprehensive, but not as good as ‘A’ and ‘B’.
4. Label ‘D’: the reviews are incomprehensive, irrelevant, or trash reviews (such as advertisements).

Then two annotations for each review, from each team, is combined into the final annotation as follows: in each original annotation, an ‘A’ is counted as 3 points, ‘B’ 2 points, ‘C’ 1 point and ‘D’ 0 point. A review with a sum of 6 or 5 points gets a final label ‘A’, a review with 4 or 3 points gets ‘B’, 2 or 1 ‘C’ and 0 ‘D’.

In our system, reviews are ranked according to equation (7), and those ranked first in their test sets by our system are called “top 1 reviews”, so there are 14 “top 1 reviews”. Table 1 evaluates how all top 1 reviews coincide with the annotations. The last four columns show the number of top 1 reviews are labeled with ‘A’, ‘B’, ‘C’ and ‘D’ in human annotation, respectively. The last row is the numbers of ‘A’, ‘B’, ‘C’ and ‘D’ labels, respectively, in annotation. Obviously, most top 1 reviews are labeled with ‘A’s or ‘B’s, and no ‘D’s.

Then top N reviews are checked to see whether the reviews ranked by information distance coincide with their comprehensiveness. Table 2 shows the results:

Firstly, it can be seen from first five columns that, in 70 reviews ranked top 5 in 14 test sets, there are 14 ‘A’s, 34 ‘B’s, and 18 ‘C’s, only 4 ‘D’s. The last row shows the total number of reviews that are labeled by ‘A’, ‘B’, ‘C’ and ‘D’, respectively. More than half ‘A’ reviews are in top 2 of its test set, and almost all ‘A’s are in top 10.

Then two manual annotations are used as baselines to be compared with the result of our system, called “Baseline 1” and “Baseline 2” in Table 2. As human annotations only have four labels: ‘A’, ‘B’, ‘C’ and ‘D’, a simple method is used to change them into numeric rankings: firstly, reviews labeled ‘A’ are ranked highest, followed by ‘B’, and then ‘C’. ‘D’s

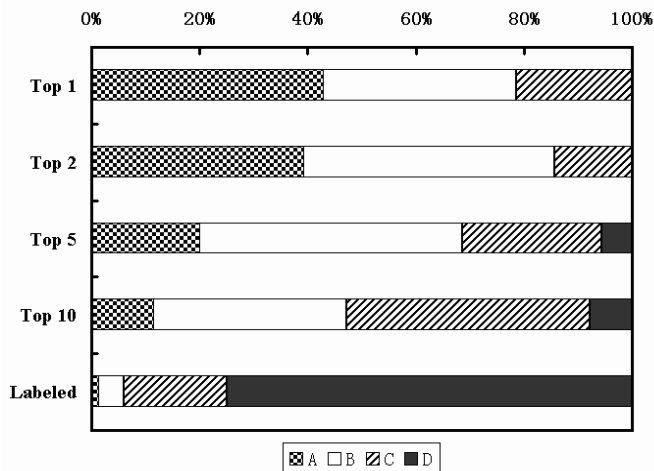


Figure 1: Proportions of ‘A’ to ‘D’ in top N results of our system and the annotation.

are all in the last. Second, reviews with the same label are ranked randomly. For most test sets, each of them has only one review labeled ‘A’, therefore, highest ranking reviews are much less affected by random. These two baselines are also measured by final annotation. The results are shown in the last eight columns of Table 2. From this table we can see human annotations differ slightly from each other. Our system result is close to manual ones.

Figure 1 shows the proportions of ‘A’, ‘B’, ‘C’ and ‘D’ in top N results of our system and in the annotation. It can be seen from the figure that in the annotation, reviews labeled ‘A’ take a fairly small part (only 1.23%), and ‘D’s take almost three quarters of the test set, while in Top 1 and Top 2 results, ‘A’s are around 40%.

5.2 Specialized Review Selection

A set containing 339 DC reviews are used as the specialized review selection task test set, and twelve features are selected. Specialized reviews selected by our system are compared with the human annotation, selected according to the most frequent opinion on a specific feature. Nine of twelve reviews selected by our system agree with human-annotated popular opinion on particular feature.

Table 3 is the result of specialized review selection. The first column contains DC features, and the second column is the most frequently used sentence or phrase to describe corresponding features, selected by human. The last column shows the relevant sentences or phrases, selected by human, of the specialized reviews selected by our system. According to the table, except features “exposure”, “flash” and “memory”, reviews selected by our system agree with the humanly selected opinion very well.

6. CONCLUSION AND FUTURE WORK

We have initially developed the theory of information distance among many objects and solved several theoretical problems. We have provided a framework so that such a theory can be applied to review mining. We have actually built a comprehensive and specialized review mining system

Table 3: Specialized Review Selection. Third column: the reviews are automatically selected, but the sentence / phrase for a specific feature is manually picked.

Feature	Popular Opinion of the Review Set	Sentence or Phrase of Selected Review
battery	battery life is long/good/great	... battery life is exceptional ...
exposure	missing manually adjustable exposure	... for low lighting conditions increase the exposure time and you are done ...
flash	flash is enough/good	... the flash does not go far enough ...
image	good pictures	... great pictures ...
lens	lens error	... lens retraction problem ...
memory	not enough/need an extra memory card	... don’t even really need a larger memory card ...
pixel	6 megapixels	... 6 megapixels ...
price	good price	... excellent price ...
screen	large screen	... large bright screen ...
shutter	fast/quick shutter	... fast shutter lag ...
video	high quality	... has good resolution even with videos ...
zoom	low optical zoom	... a higher optical zoom would have been appreciated ...

based on our new theory and demonstrated the performance of our system.

In the future work, we will further improve our approach by using sentiment classification method. For example, while talking about a type of digital camera, “great battery life” has the same meaning with “long battery life”, therefore, if sentiment information is considered, reviews selected by our improved approach may be more typical and the accuracy can be promoted.

Acknowledgment

The work of CL was done at the University of Waterloo as an exchange student, financially supported by China State-funded Study Abroad Program. ML is supported by NSERC RGPIN46506 and a Canada Research Chair program; BM is supported by NSERC and a Canada Research Chair program.

7. REFERENCES

- [1] C. Ané and M. Sanderson. Missing the forest for the trees: Phylogenetic compression and its implications for inferring complex evolutionary histories. *Systematic Biology*, 54(1):146–157, 2005.
- [2] T. Arbuckle, A. Balaban, D. Peters, and M. Lawford. Software documents: Comparison and measurement. In *The Nineteenth International Conference on Software Engineering and Knowledge Engineering*, July 2007.
- [3] D. Benedetto, E. Caglioti, and V. Loreto. Language

Table 2: Top N Comprehensive Reviews

TOP	Our Approach				Baseline 1				Baseline 2			
	A	B	C	D	A	B	C	D	A	B	C	D
1	6	5	3	0	7	7	0	0	8	6	0	0
2	11	13	4	0	11	15	2	0	13	11	4	0
5	14	34	18	4	15	34	21	0	15	24	31	0
10	16	50	63	11	17	54	69	0	16	43	81	0
Labeled	17	66	264	1034	17	66	264	1034	17	66	264	1034

- trees and zipping. *Physical Review Letters*, 88(4):048702, 2002.
- [4] C. Bennett, P. Gacs, M. Li, P. Vitányi, and W. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, July 1998.
- [5] C. Bennett, M. Li, and B. Ma. Chain letters and evolutionary histories. *Scientific American*, 288(6):76–81, June 2003.
- [6] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, August 1998.
- [7] P. Chaovalit and L. Zhou. Movie review mining: a comparison between supervised and unsupervised classification approaches. In *HICSS*, page 112c, January 2005.
- [8] X. Chen, B. Francia, M. Li, B. Mckinnon, and A. Seker. Shared information and program plagiarism detection. *IEEE Transactions on Information Theory*, 50(7):1545–1550, July 2004.
- [9] A. Chernov, A. Muchnik, A. Romashchenko, A. Shen, and N. Vereshchagin. Upper semi-lattice of binary strings with the relation “x is simple conditional to y”. *Theoretical Computer Science*, 271:69–95, 2002.
- [10] R. Cilibrasi and P. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.
- [11] R. Cilibrasi, P. Vitányi, and R. de Wolf. Algorithmic clustering of music based on string compression. *Comput. Music J.*, 28(4):49–67, 2004.
- [12] R. L. Cilibrasi and P. M. Vitányi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, March 2007.
- [13] M. C. de Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *The fifth international conference on Language Resources and Evaluation (LREC)*, May 2006.
- [14] K. Emanuel, S. Ravela, E. Vivant, and C. Risi. A combined statistical-deterministic approach of hurricane risk assessment. *manuscript, Program in Atmospheres, Oceans, and Climate, MIT*, 2005.
- [15] M. Gamon, A. Aue, S. C. Oliver, and E. Ringger. Pulse: Mining customer opinions from free text. In *International Symposium on Intelligent Data Analysis (IDA)*, pages 121–132, October 2005.
- [16] M. Hayashida and T. Akutsu. Image compression-based approach to measuring the similarity of protein structures. In *The 6th Asia-Pacific Bioinformatics Conference*, pages 221–230, 2008.
- [17] M. Hu and B. Liu. Mining and summarizing customer reviews. In *The 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, August 2004.
- [18] E. Keogh, S. Lonardi, and C. Ratanamahatana. Towards parameter-free data mining. In *The 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 206–215, August 2004.
- [19] S. Kirk and S. Jenkins. Information theory-based software metrics and obfuscation. *Journal of Systems and Software*, 72:179–186, 2004.
- [20] A. Kocsor, A. Kertesz-Farkas, L. Kajan, and S. Pongor. Application of compression-based distance measures to protein sequence classification: a methodology study. *Bioinformatics*, 22(4):407–412, 2006.
- [21] A. Kraskov, H. Stogbauer, R. Andrzejak, and P. Grassberger. Hierarchical clustering using mutual information. *Europhys. Lett*, 70(2):278–284, 2005.
- [22] N. Krasnogor and D. Pelta. Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics*, 20(7):1015–1021, 2004.
- [23] M. Li, J. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–154, 2001.
- [24] M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.
- [25] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications(2nd Edition)*. Springer-Verlag, 1997.
- [26] A. Muchnik. Conditional complexity and codes. *Theoretical Computer Science*, 271(1):97–109, 2002.
- [27] A. Muchnik and N. Vereshchagin. Logical operations and kolmogorov complexity ii. In *16th Annual IEEE Conference on Computational Complexity*, pages 256–265, June 2001.
- [28] M. Nykter, N. Price, M. Aldana, S. Ramsey, S. Kauffman, L. Hood, O. Yli-Harja, and I. Shmulevich. Gene expression dynamic in the macrophage exhibit criticality. *PNAS*, 105(6):1897–1900, 2008.
- [29] M. Nykter, N. Price, A. Larjo, T. Aho, S. Kauffman, O. Yli-Harja, and I. Shmulevich. Critical networks exhibit maximal information diversity in structure-dynamics relationships. *Physical Review Letters*, 100:058702 (1–4), 2008.

- [30] H. Otu and K. Sayood. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, 19(6):2122–2130, 2003.
- [31] H. Pao and J. Case. Computing entropy for ortholog detection. In *International Conference on Computational Intelligence*, December 2004.
- [32] D. Parry. Use of Kolmogorov distance identification of web page authorship, topic and domain. In *Workshop on Open Source Web Inf. Retrieval*, 2005.
- [33] A. M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 339–346, October 2005.
- [34] S. Rahmati and J. Glasgow. Noise tolerance of universal similarity metric applied to protein contact maps comparison in two dimensions. *manuscript, Queen Univ*, 2008.
- [35] C. Santos, J. Bernardes, P. Vitányi, and L. Antunes. Clustering fetal heart rate tracings by compression. In *The 19th IEEE Symposium on Computer-Based Medical Systems*, June 2006.
- [36] A. Shen and N. Vereshchagin. Logical operations and kolmogorov complexity. *Theoretical Computer Science*, 271:125–129, 2002.
- [37] A. Siebes and Z. Struzik. Complex data: Mining using patterns. *Lecture Notes in Computer Science*, 2447:205–229, 2002.
- [38] W. Taha, S. Crosby, and K. Swadi. A new approach to data mining for software design. *manuscript, Rice Univ*, 2006.
- [39] P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 32–44, July 2002.
- [40] N. Vereshchagin and M. V’yugin. Independent minimum length programs to translate between given strings. *Theoretical Computer Science*, 271:131–143, 2002.
- [41] M. V’yugin. Information distance and conditional complexities. *Theoretical Computer Science*, 271:145–150, 2002.
- [42] X. Zhang, Y. Hao, X. Zhu, and M. Li. Information distance from a question to an answer. In *The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2007.
- [43] L. Zhuang, F. Jing, and X. Zhu. Movie review mining and summarization. In *ACM 17th Conference on Information and Knowledge Management (CIKM)*, pages 43–50, November 2006.