Associative Feature Selection for Text Mining

Tien Dung Do, Siu Cheung Hui and Alvis C.M. Fong

Nanyang Technological University, School of
Computer Engineering, Singapore 639798

{pa0001852a, asschui, ascmfong}@ntu.edu.sg

Abstract

With the exponential growth of the number of documents available on the Internet, automatic feature selection approaches are increasingly important for the preprocessing of textual documents for data mining. Feature selection, which focuses on identifying relevant data, can help reduce the workload of processing huge amounts of data as well as increase the accuracy for the subsequent data mining tasks. In this paper, we propose a new feature selection approach for text mining based on association rules. An evaluation on the performance of the proposed associative feature selection approach based on a dataset of published data mining papers is also presented.

Keywords: Feature selection, association rules, text mining.

I. Introduction

Text mining applications always need to deal with large and complex datasets of textual documents that contain much irrelevant and noisy information. Feature selection aims to remove that irrelevant and noisy information by focusing only on relevant and informative data for use in text mining. Feature selection can be supervised with human support in labeling the data, or be unsupervised without any human involvement. In supervised feature selection, a labeled training set is first trained to derive the model, which is then used to predict an unlabelled test set [1]. To obtain a good prediction from the test dataset, it requires identifying the similarities between the training and test datasets, which makes this approach inflexible. Unsupervised feature selection does not need a pre-labeled dataset. Instead, heuristics are used for estimating the quality of the features [2]. As such, it saves cost and time on labeling the data and avoids the problem on inaccuracy of homogeneity between the training and test datasets as in the supervised process.

In this paper, we propose an unsupervised feature selection approach for text mining. The approach is based on an assumption that relevant features in a textual document dataset are closely associated. Highly associated features are searched using association rule mining [3]. In this paper, we also introduce a measure for association rule mining called *relative confidence* [4], which can truthfully reflect the relations of the items that can be used to improve the quality of the selected features. An experiment has been conducted to evaluate the performance of the proposed associative feature selection approach.

The rest of the paper is organized as follows. In the next section, we review the concepts on association rules and introduce the *relative confidence* measure. Section 3 presents the related work on feature selection and our proposed approach for feature selection using association rule mining. Section 4 describes an experiment that evaluates the performance of the proposed associative feature selection approach. Section 5 concludes the paper.

II. Association Rules

Formally, an association rule *R* is an implication $X \Rightarrow Y$, where *X* and *Y* are sets of items in a given dataset [3]. The *confidence* of the rule *conf(R)* is the percentage of transactions that contains *Y* amongst the transactions containing *X*. The *support* of the rule *supp(R)* is the percentage of transactions containing *X* and *Y* with respect to the number of all transactions.

Let P[S] be the probability of an itemset *S* present in a certain transaction of the database. Similar to the definition of *support* of a rule, P[S] can be considered as the *support* of the itemset S (it is denoted as *supp(S)* in some papers). We call P[X] the *antecedence support* and P[Y] the *consequence support* of the rule *R*. Assume that the database contains *N* transactions with the numbers of transactions that contain *X*, *Y*, and both *X* and *Y* are *a*, *b*, and *c* respectively. It can be implied from the definitions of *support* and *confidence* of association rules that *supp(R) = c/N* and *conf(R) = c/a*; and from the definition of *support* of an itemset that P[X] = *a/N*, P[Y] = *b/N* and $P[X \wedge Y] = c/N$. Thus, the values of *supp(R)* and *conf(R)* can be computed using $P[X \wedge Y]$ and P[Y] as follows:

$$\text{supp}(R) = P[X \wedge Y] \quad (1)$$

$$\text{conf}(R) = \frac{P[X \wedge Y]}{P[X]} \quad (2)$$

The *confidence* of a rule *R* measures the implication relation of the antecedence (*X*) to the consequence (*Y*), which is the actual interestingness to the rule. It shows the prediction of *Y* when *X* occurs. For example in a market basket analysis problem, in 80% of the cases when people buy bread, they also buy milk. The high value of *confidence* means a high probability of implication from bread to milk. This may suggest that the supermarket should put milk next to bread so that it is more convenient to most of the customers buying bread who (80%) will then look for milk.

The drawback of the *confidence* measure is that it is purely an absolute value of the implication of the rule "*X* to *Y*". This implication is affected not only by the relation of *X* to *Y*, but also by the distribution of *Y* (the *consequence support* of the rule). When the *support* value of *Y* is high, *Y* will likely be present in any transactions including those containing *X* regardless of how related *X* is to *Y*. In the next section, we will introduce a measure for association rules called *relative confidence* [4], which can truthfully reflect the relations of the items. The measure, therefore, can be used to improve the quality of association rules.

A. Relative Confidence

The *confidence* of a rule *R* reflects the implication "if a transaction contains *X*, then it will probably contain *Y*". The transaction will contain *Y* because *X* is related to *Y* so that once it contains *X* it will be likely to contain *Y*, or because *Y* is contained in some transactions of the database so that it will possibly be contained in the transaction. Therefore, the *confidence* of *R* is influenced by two factors:

(i) the degree on how *X* is related to *Y*; and (ii) the density of random distribution of transactions containing *Y* amongst transactions containing *X*. The first factor can be considered as the actual value of the relationship from *X* to *Y*. We call this as *relative confidence* of the rule *R* (denoted as *Rconf*(*R*)). The second factor is the same as the degree of distribution of the transactions containing *Y* on the whole database of transactions which is represented by the *support* of *Y* (i.e. *P*[*Y*]).

Let *A* be the set of transactions containing *X* and *B* be the set of transactions in *A* that contains *Y*, then *conf(R) = p(B|A)* (probability of *B* on *A*). According to the two factors on the *confidence* of the rule *R*, *B* would be compounded from (i) a set of transactions B_1, which reflects the relationships of *X* to *Y*; and (ii) a set of transactions B_2, which reflects the random distribution of the transactions containing *Y*. We can now imply that $p(B_1|A) = Rconf(R)$, $p(B_2|A) = P[Y]$ and $B = B_1 \cup B_2$. We can then find the relationship of the probabilities of these sets of transactions. In the following equations, the probabilities are considered in the sample space of *A*. It means that, for example, *p(B)* denotes *p(B|A)*.

$$\begin{aligned} p(B) &= p(B_1 \cup B_2) \\ &= p(B_1) + p(B_2) - p(B_1 \cap B_2) && (3) \\ &= p(B_1) + p(B_2) - p(B_1) \times p(B_2) && (4) \\ p(B_1) &= \frac{p(B) - p(B_2)}{1 - p(B_2)} && (5) \end{aligned}$$

Equations (3) and (4) are based on probability theorems of union of arbitrary events and multiplication rule for independent events [5]. Note that the two sets B_1 and B_2 are independent because B_2 is a result from a random distribution. Now, we replace the equations on the probabilities of *B*, B_1 and B_2 into (5) to derive the formula for the *relative confidence* of *R*:

$$\begin{aligned} Rconf(R) &= \frac{conf(R) - P[Y]}{1 - P[Y]} \\ &= \frac{\frac{P[X \wedge Y]}{P[X]} - P[Y]}{1 - P[Y]} \\ &= \frac{P[X \wedge Y] - P[X] \times P[Y]}{P[X] - P[X] \times P[Y]} \end{aligned} \qquad (6)$$

Definition 1. The *relative confidence* of a rule *R* ($X \Rightarrow Y$) is defined as:

$$Rconf(R) = \frac{P[X \wedge Y] - P[X] \times P[Y]}{P[X] - P[X] \times P[Y]} \qquad (7)$$

III. Associative Feature Selection

Feature selection is used to select a "better" subset that can describe data from an original dataset. The aims of feature selection are to (i) focus on the relevant data; and (ii) reduce the amount of data [2]. Feature selection approaches are based on either exhaustive or heuristic search. Exhaustive feature selection approaches search for all possible combinations of features and find an optimal one based on an evaluation criterion. Let *N* denote the number of features in the original dataset. The total number of candidate subsets is 2^N. Although it is not always necessary to scan all possible

subsets, exhaustive search is still quite computationally expensive [6]. Heuristic feature selection approaches employ heuristics in conducting search. One approach of heuristic search is to score features based on some heuristic measures. The score of a feature indicates how relevant it is to the dataset. The selection process is then very straightforward. Features will be selected if their scores are above a predefined threshold.

A. Feature Selection for Text Mining

Recently, there is a tremendous growth in the number of documents available on the Internet. These documents with unstructured data have become the predominant data type stored online. Text mining [7], therefore, is one of the most important tasks in data mining. The "bag-of-words" approach is commonly used to analyze textual documents. In this approach, a document is considered as a set of words or phrases (called terms). When applying data mining techniques to textual documents, a document is considered as an instance (or transaction), while terms (words or phrases) are considered as features (or items).

There are a number of feature selection approaches, which can be applied effectively to textual data. Most of them are based on a scoring scheme of terms [8]. The score of features represents the quality of the terms in the document dataset. A term with high score means it is important or relevant to the dataset. In supervised approaches, term scores are based on labeled training set, which is class information. Some of the popular supervised feature selection approaches are information gain (IG), mutual information (MI) and χ^2 statistics (CHI) [9].

Unsupervised feature selection approaches are based on heuristics for estimating the quality of terms in a dataset. For a dataset of textual documents, the heuristics generally focus on term distribution among the dataset. The popular unsupervised feature selection approaches include document frequency (DF) and term strength (TS). Document frequency (DF) is a simple but effective measure for feature selection. Yang *et al.* [9] concluded that DF is among the best measures (as good as IG and CHI) for selecting informative features. Document frequency of a term is the number of documents in which the term occurs. The feature selection approach calculates document frequency for every term and removes the terms whose document frequency is less than a predefined threshold. The basic assumption is that frequent terms are more important and relevant to the dataset in comparison to the infrequent ones.

Term strength is proposed in [10] initially for stop-word removal. This approach estimates the strength of a term based on how likely it appears in "closely-related" documents. It is based on a heuristic that documents with many shared words are related, and that terms in heavily overlapping area of related documents are relatively informative [9]. The approach has two steps:

- *Finding pairs of similar documents.* This step calculates the similarities between all pairs of documents in the dataset $sim(d_i,d_j)$ using the cosine value of the two document vectors. Two documents d_i and d_j are then considered "similar" if $sim(d_i,d_j)$ is above a predefined threshold ξ.
- *Calculating term strength.* Strength of a term $s(t)$ is computed based on the estimated conditional probability that the term t occurs in a document d_i when it occurs in document d_j, which is similar to d_i: $s(t) = p(t \in d_i \mid t \in d_j \wedge sim(d_i, d_j) \geq \xi)$.

As we have mentioned earlier, unsupervised feature selection approaches save the cost of labeling data and avoids the problem on inaccuracy of homogeneity between training and test datasets in the supervised process. This characteristic is especially important for text mining tasks in which we always need to deal with a huge amount of documents of various topics. In the next section, we will

propose a new approach for unsupervised feature selection using association rules. Unsupervised approaches such as DF and TS will also be implemented for comparison with the performance of the proposed associative feature selection approach.

B. Feature Selection using Association Rules

In the previous sections, we have shown how terms could be scored based on their distribution on the document dataset. The distribution of a term could be examined independently like DF or in relation with other terms like TS. In this section, we propose an approach to select relevant terms based on the associations among them. Such associations can be discovered using association rule mining.

Generally, a document of a dataset belongs to one or more *topics* in a certain field or area. The topics of all documents in the dataset form the theme for the dataset called a *domain*. In short, a domain can contain multiple topics and each topic is discussed in some documents. A term that is relevant to the dataset means it is relevant to the domain of the dataset. A relevant term should then be relevant to some topics of the domain. It may be used to explain or illustrate the topics or some concepts of the topics. As terms relevant to a topic will possibly occur in documents belonging to the topic, a relevant term is likely to occur with some other relevant terms (which are related to the same topics with it). The remaining irrelevant terms, which are not related to any topics, are probably distributed randomly. The probability of these terms to occur in a document does not depend on the topics of the document. This is illustrated in Figure 1. The associative features of relevant terms and irrelevant terms from documents in a domain can be summarized as follows:

- A relevant term is probably associated with other relevant terms.
- An irrelevant term is not likely to be associated with other terms.

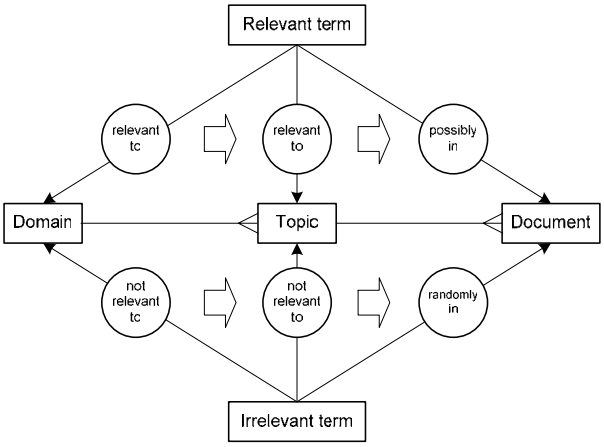

Figure 1. Relations of relevant and irrelevant terms in a domain.

The proposed associative feature selection approach is based on the heuristics discussed above to separate relevant and irrelevant terms. The occurrence of terms in many association rules means that they are associated with many other terms. These terms should then be assigned with a high score so that they are considered as relevant terms. On the contrary, terms that occur infrequently in association rules should be assigned with a low score of irrelevant terms. The assignment of scores to features comprises the following three steps: (1) we first determine the constraints of the association rules; (2) we search for association rules satisfying the constraints; and (3) the features are then scored based on the association rules found.

In the first step, we determine the constraint *F* for the association rules: *F: ARs* $\rightarrow$ *Boolean*. Here, *ARs* is the set of all association rules on the set of terms. Conventionally, the constraint for association rules is that *support* and *confidence* are greater than the minimal values (thresholds) of *min_supp* and *min_conf*. The constraint *F* can also include other measures for mining association rules. In our implementation, the *relative confidence* measure will be used to improve the quality of the selected terms.

The second step searches for association rules that satisfy the constraint *F*. The typical approach for searching association rules is the Apriori algorithm given in [3]. In this approach, the *support* measure is first used to filter out most of the rules that do not satisfy *min_supp*, the *confidence* measure (or other measures) is then applied to the remaining rules that also satisfy *min_conf*.

Finally, terms are scored based on the association rules found in the second step. A term will have a high score if it appears in many rules. The score calculation of the term can also be based on the "quality" of the rules (for example, support and confidence), position of the term (antecedence or consequence) and the scores of other terms in the set of rules it occurs.

IV. Performance Evaluation

An experiment has been conducted to measure the performance of the proposed associative feature selection approach. A document dataset comprising 514 published data mining papers was downloaded from searching the Science Direct web page (http://www.sciencedirect.com) on June 2003 with the searching keyword of "data mining" in the title, abstract and keyword fields. The size of the abstract documents varies from 2 to 27 lines of text. The dataset is obtained from the data mining papers with the intention that the terms found under the measures could be evaluated by their meanings, that are familiar to the authors of this paper. For example, we can recognize easily that "neural" and "network" are relevant to data mining, while "university" and "student" may not.

The preprocessing of the document dataset consists of three steps. First, a stop-list of 571 common words [11] is used to eliminate the stop-words from all documents. Second, a stemming process of suffix removal is carried out to generate word stems. As a result, 622 terms of which document frequencies greater than or equal to 10 (2% of the 514 documents) are selected from a total of 5377 words generated from the second step.

We have implemented the associative feature selection approach and other approaches using document frequency (DF) and term strength (TS) for comparison. All approaches are based on a term scoring scheme. The DF approach scores terms based on frequency. The frequency of a term is the number of documents containing the term. The TS approach scores terms based on their strength [10]. It first calculates the similarities *sim(d,e)* of all pairs of documents in the dataset. Pairs of documents are then selected if the similarities are above a predefined threshold. Finally, the strength of terms is estimated based on these pairs of documents. In associative feature selection, we use the scoring procedure discussed in Section 3.2 to calculate the scores. In this approach, there are two implementations: one uses the *confidence* measure (called AR) and the other uses the *relative confidence* measure (called R-AR). The threshold values for *support* and *confidence* are 2% and 30% respectively.

Selection of terms after scoring is straightforward. A threshold is then defined and the features that have scores above the threshold value are selected. The threshold can be determined by a statistical

approach as in [10]. The number of terms can be altered by increasing or decreasing the threshold value.

A. Measuring Goodness of Features

The implementation of different approaches would generate different sets of relevant and irrelevant terms from the total of 622 terms selected. A standard labeling (relevant or irrelevant) of terms, therefore, is needed so that we can evaluate the goodness of each feature selection approach. In this experiment, the labeling is obtained in two steps. In the first step, terms are manually classified into five groups as follows:

- *Topics, tasks, approaches, applications:* association, classify, cluster.
- *Concepts, terms:* term, pattern, set, database, text, algorithm. They are concepts used more frequently in data mining than other IT topic.
- Words specially used for a *topic* of data mining: frequency, large itemset, apriori (association rule mining); gene, tissue (data mining for biology); attribute, dimension (database mining); sequence, parallel, regression (data mining approaches).
- Words that are also *popular* in other IT topics: system, approach, software.
- *Common* words: show, define, increase, analyze, accurate, automatic, intelligent, challenge.

The second step simply groups the three first groups into a set labeled as relevant terms. This set consists of a total of 118 terms. The terms in groups 4 and 5 which are also used in documents in domains other than "data mining" are labeled as irrelevant. Here, the classification of terms into the above five groups makes the "manual labeling" process easier and more accurate. It also gives the justification on which items are relevant or irrelevant. The evaluation uses the standard *precision* and *recall* measures based on the above document dataset. The *precision* and *recall* measures [12] are calculated as follows:

$$precision = \frac{terms\ selected\ and\ relevant}{total\ terms\ selected}$$

$$recall = \frac{terms\ selected\ and\ relevant}{total\ terms\ relevant}$$

B. Results

The results of each implementation are a list of terms with the corresponding scores. To select k relevant terms, we simply sort the list of terms based on the scores and obtain the k terms with the highest scores. To evaluate the goodness of each implementation, the k relevant terms are selected and compared to the standard labeled set using the precision and recall measures. We have conducted the experiment with a wide range of k for each implementation. The range of k has been set from 59 (a half of the number of relevant terms of 118) to 236 (two times of 118). The results are shown in Figure 2.

The experimental results have shown that the associative feature selection approach is better than the approach using document frequency and as good as when the term strength is used. It can be explained that document frequency while focusing only on frequent terms may miss some important but rare ones. Association rules (AR) weighting is similar to term strength (TS) in considering the relations among terms. These two approaches are based on the heuristic that relevant terms on a textual document dataset should have some relations on the distribution among the dataset.

The experiment has also shown that the new measure *relative confidence* of association rules is appropriate for discovering good rules. The precision and recall measures have increased considerably when the *relative confidence* measure is used instead of the *confidence* measure.

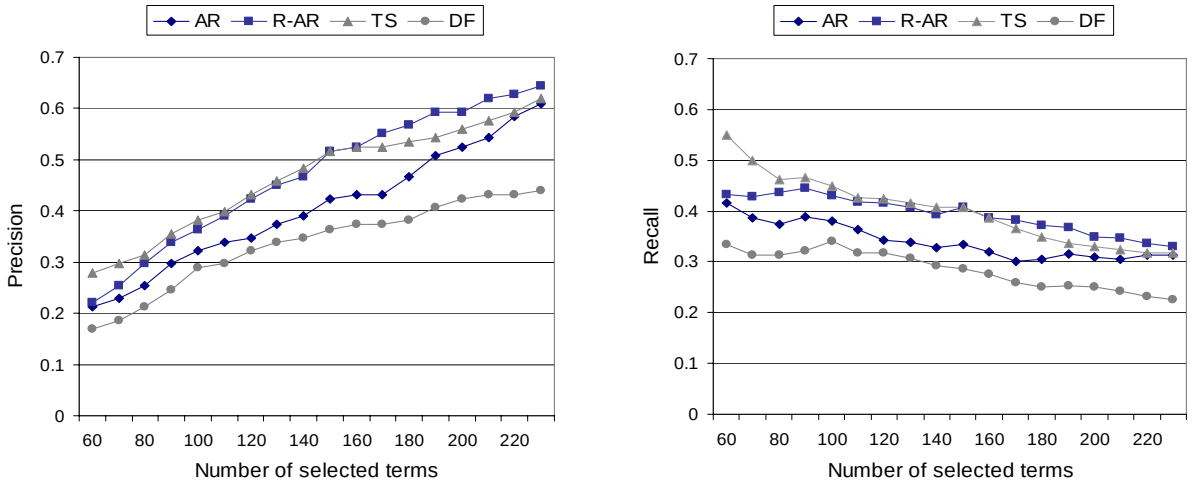

Figure 2. Precision and recall values of the different feature selection approaches.

V. Conclusion

In this paper, we have proposed an unsupervised associative feature selection approach for text mining. Based on association rule mining, the approach discovers the association strength of terms among documents. This hidden knowledge is then used to score the importance of the terms in the dataset automatically. This paper has also introduced the *relative confidence* measure, which aims to truthfully reflect the relation of items by removing the noisy substance of frequency from the *confidence* measure, for mining association rules. This measure is especially important when mining datasets of terms with high frequency such as textual documents.

The unsupervised feature selection approaches using TS and AR exploit the co-occurrence of terms on textual documents. However, with the examination of all the pairs of documents on the dataset, the time complexity of the approach using TS would be $O(N^2)$ where N is the number of documents. This makes the approach a weak choice for large datasets of documents. The time performance of the approach based on AR depends on the process of searching for association rules. Some AR mining approaches have very good time performance, which is directly proportional to the size of the dataset.

In addition, the proposed associative feature selection approach using association rule mining is only evaluated based on the criterion of how relevant the selected features are to the document dataset. To have a more accurate evaluation of the approach, further research should be performed on its effectiveness to the later data mining process such as categorization. For example, Yang *et al.* [9] have examined different feature selection approaches for text categorization. Several feature selection approaches are implemented and the features selected by different approaches are then used for categorization purposes. The evaluation of the categorization process could then be used to evaluate the effectiveness of different feature selection approaches for text mining.

References

[1] S. Wu, P.A. Flach, *Feature Selection with Labelled and Unlabelled Data.* In Proc. of ECML/PKDD'02 workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning, University of Helsinki (2002) 156-167.

[2] H. Liu, M. Motoda, L. Yu, *Feature Extraction, Selection, and Construction.* In N. Ye (eds.): The Handbook of Data Mining, Lawrence Erlbaum Associates, Inc. Publishers (2003) 409-423.

[3] R. Agrawal, R. Srikant, *Fast algorithms for mining association rules.* In Proc. of the 20th Int'l Conf. on Very Large Databases (VLDB '94), Santiago, Chile (1994) 487-499.

[4] T.D. Do, S.C. Hui, A.C.M. Fong, *Mining Association Rules with Relative Confidence.* In Proc. of the Fifth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'04), Exeter, UK (2004) 306-313.

[5] E.K. Szig, *Advanced Engineering Mathematics.* 7th edn. Johl Wiley & Sons Inc (1993) 1155-1158.

[6] D. Koller, M. Sahami, M, *Toward optimal feature selection.* In S. Francisco (eds.): Thirteenth Interna-tional Conference on Machine Learning (ICML '96), Lorenza Saitta, (1996) 284-292.

[7] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms.* Wiley-IEEE Press (2003).

[8] D. Mladenic, *Feature subset selection in text-learning.* In Proc. of European Conference on Machine Learning (1998) 95-100.

[9] Y. Yang, J.O. Pedersen, *A comparative study on feature selection in text categorization.* In Proc. of the 14th International Conference on Machine Learning (ICML-97), Morgan Kaufmann Publishers, San Francisco, US (1997) 412-420.

[10] J.W. Wilbur, K. Sirotkin, *The automatic identification of stop words.* J. Inf. Sci, Vol. 18 (1992) 45-55.

[11] WordNet Project of Princeton University. Available online at <http://wordnet.princeton.edu>

[12] C. Van Rijsbergen, Information Retrieval. Utterworths, London, England (1979).

Mr. Do Tien Dung is currently a Ph.D student in the Nanyang Technological University, Singapore. He received his Bachelor of Engineering degree in Computer Science from Hanoi University of Technology, Vietnam in 1995. His research interests include data mining and Web mining.

S. C. Hui is an Associate Professor in the School of Computer Engineering at Nanyang Technological University, Singapore. His current research interests include data mining, Internet technology, and multimedia systems. Previously, he worked in IBM China / Hong Kong as a system engineer from 1987 to 1990. He received his B.Sc. degree in Mathematics in 1983 and a D. Phil degree in Computer Science in 1987 from the University of Sussex, UK. Dr. Hui is a member of IEEE and ACM.

A.C.M. Fong is currently Assistant Professor in Computer Engineering at Nanyang Technological University. His research interests include various aspects of Internet technology, information theory, and video and image signal processing. Previously, he was with the Motorola Corporate Research and Technology Center. He received his degrees from the University of Auckland and Imperial College, London. Dr. Fong is a member of IEEE and IEE, and is a Chartered Engineer.