Blog Classification Using Tags: An Empirical Study*

Aixin Sun[1], Maggy Anastasia Suryanto[1], and Ying Liu[2]

[1] Nanyang Technological University, Singapore
axsun@ntu.edu.sg
[2] Hong Kong Polytechnic University, Hong Kong, China
mfyliu@polyu.edu.hk

Abstract. With an exponential growth of Weblogs (or blogs), many blog directories have appeared to help users to locate topical blogs. As tags are commonly used to describe blogs, we study the effectiveness of tags in blog classification. Compared with titles and descriptions, our experiments, using 24,247 blogs, showed that tags could lead to better classification accuracy. It is interesting to observe that more tags did not necessarily lead to better classification accuracy. To better describe blogs, we have also proposed a tag expansion algorithm that assigns a blog more tags that are often co-occur with those already associated with the blog. Our experiments showed that tag expansion helped to improve the recall of blog classification with the price of precision degradation.

1 Introduction

Blogs are online personal diaries managed by software packages that allow single-click publishing [5]. Each diary entry in a blog is also known as a blog post (or post). These blog posts are often displayed in reverse chronological order and their contents include personal views, observations, discussions, and other topics. The rapid growth of blogs has created new research opportunities in information retrieval, text mining, social studies, and many other areas.

Similar to Yahoo! Directory organizing Web sites/pages into topical categories, a number of blog directories are now available online. Examples are BlogFlux[1] to classify blogs into 161 flat topical categories; BlogCatalog[2] to organize blogs into hierarchical topical categories with 49 top-level categories; and BOTW[3] to list blogs in a hierarchy with 12 top-level categories. These blog directories provide an easy way of locating blogs of certain topic(s), in addition to blog searching. While many of these blog directories require manual assessment

* This research is supported by grant SUG7/06, Nanyang Technological University, Singapore.

[1] http://dir.blogflux.com/, accessed on Jun 24, 2007.
[2] http://www.blogcatalog.com/, accessed on Jun 24, 2007.
[3] http://blogs.botw.org/, accessed on Jun 24, 2007.

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 307–316, 2007.

of blogs, which is labor intensive and time consuming, automatic blog classification methods offer an attractive option. *Blog classification* refers to the task of assigning blogs one or more pre-defined categories.

The problem of blog classification is different from most text/Web classification problems because of at least three reasons. Firstly, the object to be classified in blog classification are blogs where each blog consists of a set of blog posts and the number of posts may vary significantly from one blog to another. On the other hand, the objects to be classified in text/Web classification are individual text/Web documents, e.g., news articles [12]. Secondly, by its nature, a blog is frequently updated with newly published blog posts making blogs, the objects to be classified, rather dynamic compared with those documents involved in text/Web classification. Thirdly, as a type of user generated content, a blogger may write any topic of his/her interests and the topics of blog posts could be very diverse. The diversity of topics and dynamic updating nature make blog classification a much more challenging task compared with text/Web classification.

A blog could be described using features derived from its various properties, such as, title, description, tags, blog posts and so on. Among them, tags represent a new type of user-generated data that is not available for most text/Web documents. Although tags have been receiving much interests from researchers in various areas (see Section 2), the impact of using tags for blog classification has not been studied, to the best of our knowledge. On one hand, tags are often considered as "metadata" to describe the associated object (the blog in this case), which is believed to be indicative in blog classification. On the other hand, it is well known that tags are given by users without referencing to any controlled vocabulary. For this reason, different terms having similar semantics may be chosen by users to tag blogs of similar topics, and the same term maybe used to tag blogs of different topics as users may have different understanding on the scope of each topic.

In this paper, we study the effectiveness of tags in blog classification and try to answer the following three questions: (i) are tags more effective in blog classification than other type of data, e.g., title and description? (ii) is it true that more tags lead to more accurate classification? (iii) does tag expansion help in getting better classification accuracy? Tag expansion, similar to query expansion, refers to the process of expanding tags with terms having similar semantics. Tag expansion partially solves the problem of having different terms tagging blogs of the same topic.

To answer these questions, we conducted our experiments using 24,247 blogs collected from BlogFlux and classified them into 20 categories using Support Vector Machines (SVM) classifiers [8,15]. From our experimental results, we observed that tags were more effective in blog classification than features extracted from blog title and description although the latter usually contain more terms than tags. On the other hand, title and description are complementary to tags and the best classification accuracy was achieved when all these features were used together. Our experimental results surprisingly showed that more tags did not necessarily lead to better classification accuracy. To answer the third question,

we proposed a tag expansion algorithm based on Personalized PageRank algorithm [7] using co-occurrence relationships among tags. Evaluated in our experiments, it is observed that the tag expansion algorithm could improve the classification accuracy only when the blogs have relatively more tags. Such an interesting observation calls for further study on the topic, which is also a part of our future work.

This paper serves as pilot study on automated blog classification using tags. The observations obtained from our experimental results could benefit future studies in this area. The rest of the paper is organized as follows. In Section 2, we survey the related studies on tagging and blog classification. We present our data corpus in Section 3 followed by the tag expansion algorithm in Section 4. Our experimental results are presented in Section 5. Finally, we conclude the paper in Section 6.

2 Related Work

Tagging has been receiving much attention from more and more researchers in various areas such as social studies and text/Web mining. Marlow *et al.* [10] summarized tagging systems used by various web sites and presented a taxonomy of tagging systems with 7 dimensions including the three (with their main categories) shown in Table 1. In the Table, we also present the characteristics of data corpus used in our experiments (see Section 3). Note that, although we state that the object type in our study is textual, the tags we are interested in are those attached to blogs (as a group of blog posts) rather than those attached to individual blog posts. This makes our research very different from many other works on tags associated with blog posts [1,2,6,13].

Berendt and Hanser in [1] compared the performance of blog post classification using features derived from tags, title, and body; it is argued that tags are not metadata but "more content" as (i) tags have a low similarity with post body and (ii) tags together with body yielded better classification accuracy than any of them alone. Brooks and Montanez studied the effectiveness of using tags to organize blog posts into clusters [2]. They found that posts sharing the same tag have a lower similarity than those sharing the same extracted term (for each blog post, the top 3 words having the highest $tf \cdot idf$ scores were extracted). Nevertheless, their results showed that tags are useful in grouping posts into broad categories. In [6], it is observed that frequently occurring tags are usually good meta-labels of a cluster produced using content clustering.

Table 1. Example taxonomy dimensions and characteristics of our corpus

Dimension	Main categories	Our corpus
Tagging right	self-tagging, permission-based, or free-for-all	self-tagging
Tagging support	blind, suggested, viewable	blind
Object type	textual, non-textual	textual

Our work is also related to those work in Web classification where the task is to assign Web pages one or more pre-defined categories. In Web classification, features derived from title, content, hyperlinks and anchor words of Web pages have been evaluated [3,14,15]. Among the classifiers evaluated, SVM classifiers have demonstrated good classification accuracy [12]. Web classification techniques have recently been applied to blogs to detect spam blogs [9] and to label blog posts to be informative or affective articles [11].

3 Data Corpus

Our corpus is collected from BlogFlux directory in June 2007. All blogs are organized in 161 flat categories arranged in alphabetical order with *Academic* and *Zookeeping* being the first and last categories (as of June 2007). For each blog, BlogFlux provides the following metadata: *title*, *description*, *blogger*, *geographic location*, *language*, one or more *categories* assigned to the blog, and *tags* associated with the blog. We collected the metadata of all 76,997 blogs listed in the directory and among them 52,709 (or 68%) are in English. As our main focus is the use of tags in classification, from the English blogs, we selected those having at least 2 tags to form our corpus. The corpus is known as BFE (BlogFlux English) dataset, consisting of 24,247 blogs. Note that, we did not obtain the posts of all blogs in BFE dataset as those blogs are hosted on various blog sites and extracting blog posts from a large number of blog sites is a challenging task.

3.1 Tag Normalization

When submitting a blog for possible listing in BlogFlux, one may give zero to five tags to describe the blog. These tags could be either word tags (consisting of exactly one atomic word) or phrase tags (consisting of more than one word). For easy processing of tags, we performed tag normalization as in [13]. Word tags, if not stopwords, are stemmed using Porter's stemming algorithm. Phrase tags are first tokenized; the non-stopword tokens (or words) are stemmed and indexed as single-word tags. All stemmed words originally from the same phrase are sorted in alphabetic order to form a normalized phrase tag. For example, phrase tag `real estate` becomes two word tags `real`, `estat`, and one phrase tag `estat real` after normalization. After normalization, 42,798 distinct tags were obtained from BFE dataset.

Figure 1 shows the distribution of the number of tags against the blog frequency of each tag (i.e., the number of blogs associated with the tag). It is clear that a power law distribution is demonstrated with majority of tags appears with very few blogs only, while a few tags having blog frequency greater than 1000. Such an observation shows that distribution of tags attached to blogs is similar to that attached with blog posts [6].

We define *tag degree* of a blog to be the number of normalized tags attached to a blog, and *category degree* to be the number of categories assigned to the blog. Figures 2(a) and 2(b) illustrate the distribution of tag degree and category degree among blogs in BFE dataset. On average, each blog has 6.3 normalized

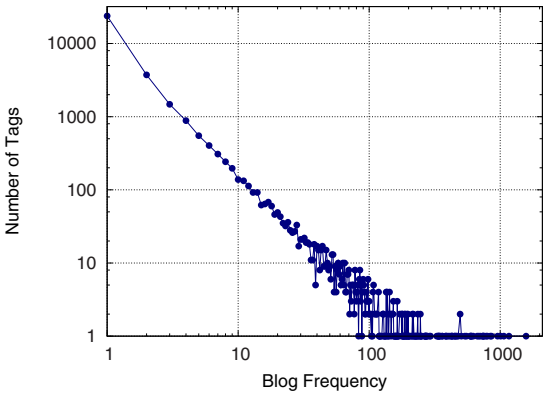
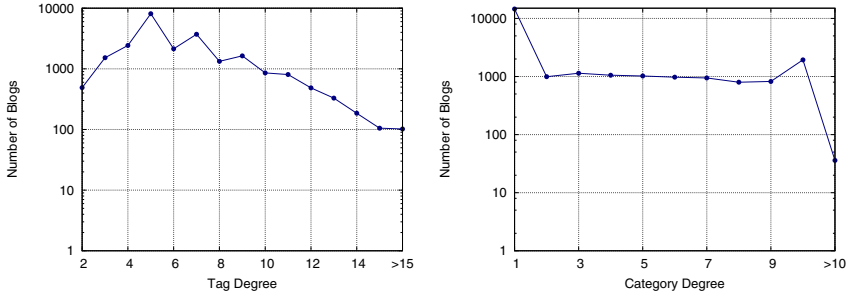


Fig. 1. Power law distribution of tags



(a) Tag Degree (b) Category Degree

Fig. 2. Tag degree and category degree distribution

tags with very few having more than 15 tags. Although most blogs are labeled to exactly one category, many of them are labeled to 2 to 10 categories with very few labeled to more than 15. On average, each blog is labeled to 3.1 categories. This suggests that topics demonstrated in blogs are often diverse and could be related to different categories.

3.2 Popular Tags and Category Names

As shown in Section 3.1, some tags are much more popular than others. We have therefore listed the 20 most popular word tags and phrase tags according to their blog frequencies, shown in Table 2. In the table, we also list the top 20 categories having most number of blogs belonging to them. It is interesting to observe that 11 of the top word tags, highlighted in bold, match the names among the top 20 categories. There are also another 12 tags, underlined, matching names among the rest 141 categories. This shows a strong relationship between the popular

Table 2. Popular word and phrase tags

Word Tag	BlogFreq
blog	1577
new(s)	1166
polit(ics)	1067
music	1008
marketing	956
art	858
travel	841
internet	800
life	784
busi(ness)	758
humor	744
technolog(y)	740
design	735
person(al)	707
web	687
photographi(y)	668
review	623
home	605
video	602
monei	595

Phrase Tag	BlogFreq
estat(e) real	249
internet market	180
design web	164
loss weight	152
make monei	110
current event	105
home work	95
engin optim search	88
cultur(e) pop	84
make monei onlin	80
busi(ness) home	78
game(s) video	78
develop(ment) web	77
financ person	69
market onlin	67
develop person	58
busi small	55
hip hop	52
affili market	52
creativ(e) write(ing)	48

Category	#Blogs
personal	4276
internet	2297
general	2262
humor	2077
entertainment	1977
computers-tech	1726
business	1681
technology	1607
art	1554
politics	1516
travel	1503
music	1411
health	1320
religion	1305
sports	1196
life	1168
photo-blog	1150
food-drink	1130
commentary	1049
opinion	995

tags and category names in blog directory, which suggests that tags could be effective features for blog classification.

4 Tag Expansion

As discussed in Section 1, different terms having similar semantics may be chosen by users to tag blogs of similar topics. To partially solve this problem, we propose a tag expansion algorithm using co-occurrence relationship among tags. The proposed tag expansion algorithm is based on the Personalized PageRank [4,7]. The basic idea is to expand the tags attached to a blog by bringing in those tags that are often used together with those former tags.

Let $\mathcal{T}$ be the set of tags. The directed graph with node set $\mathcal{T}$ and edges corresponding to co-occurrence relationships among tags is known as the *Tag Graph*. A directed edge from tag t_j to tag t_i exists if t_j and t_i are ever used together to tag one or more blogs, i.e., t_j co-occurs with t_i. Let T_b ($T_b \subseteq \mathcal{T}$) denote the set of tags attached to blog b. To expand T_b, each tag $t_i \in \mathcal{T}$ is scored using Equation 1 in an iterative manner.

$$s^{n+1}(t_i) = \alpha s^0(t_i) + (1-\alpha) \sum s^n(t_j) \times w(t_j, t_i) \tag{1}$$

In Equation 1, α is the teleportation probability and typically $\alpha = 0.15$; $s^n(t_j)$ is the score of tag t_j in the nth iteration; $s^0(t_i)$ is the initial score of t_i; and

$w(t_j, t_i)$ is the weight associated with the edge from t_j to t_i. In our experiment, $w(t_j, t_i)$ is defined by the ratio between the number of blogs tagged by both t_j and t_i and the number of blogs tagged by t_j, as shown in Equation 2 where $|C|$ refers to the number of elements in set C.

$$w(t_j, t_i) = \frac{|\{b|t_i \in T_b \wedge t_j \in T_b\}|}{|\{b|t_j \in T_b\}|} \tag{2}$$

$$s^0(t_i) = \begin{cases} \frac{tf(t_i)}{\sqrt{\sum_{t_\ell \in T_b} tf(t_\ell)^2}} & \text{if } t_i \in T_b \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The initial score of a tag t_i is given in Equation 3 where $tf(t_i)$ refers to the term frequency of tag t_i derived from all tags attached to blog b. Only those tags attached to blog b have non-zero scores; among them most of the tags have term frequency of 1. Very few tags may have term frequency more than one due to the tag normalization. For example tag `web` has term frequency of 2 if two tags are attached to a blog: `web` and `web design`.

5 Experiments

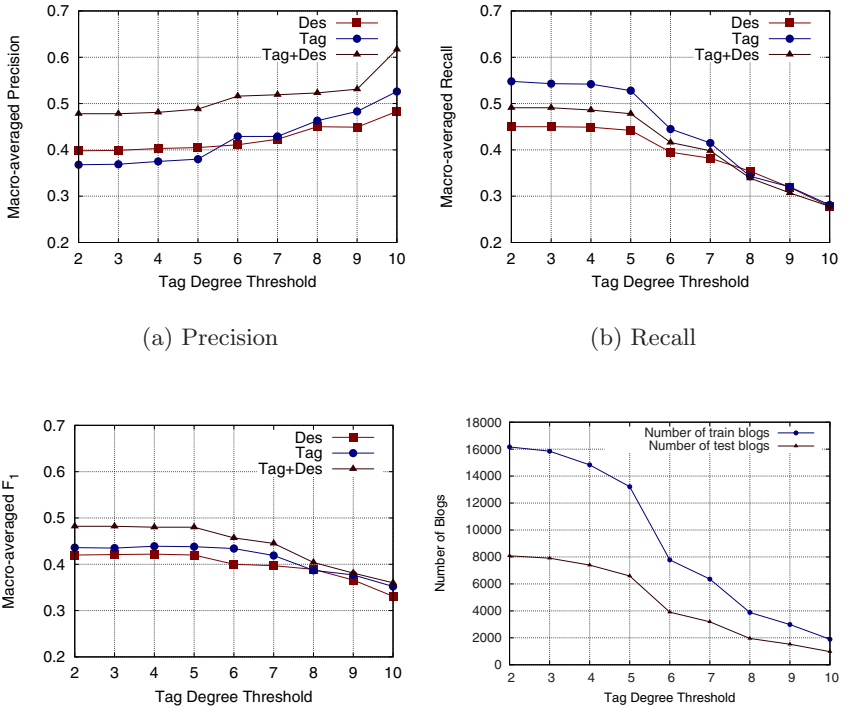
We conducted two sets of experiments. The first set is to evaluate the classification effectiveness of tags compared with other types of features derived from blogs. The second set is to evaluate the effectiveness of the proposed tag expansion algorithm in improving blog classification accuracy.

5.1 Experimental Setup

Recall that in our BFE dataset, each blog has its *title* and *description* besides its tags. We derived a text description, known as Des feature, for each blog using terms appearing in its title and description. After stopword removal and stemming, Des contains 14.8 terms on average for each blog. The number of terms in Des is about twice of the 6.3 terms contained in tags on average. In the first set of experiments, we report the classification accuracy of using features derived from tags, Des, and tags together with Des.

The BFE dataset was randomly partitioned into two sets: two-thirds blogs were used for training and the rest one-third for testing. The experiment was conducted on the top 20 categories with the largest number of blogs (see Table 2 last column) using SVM^{light} package[4]. Binary classification setting was applied. That is, one classifier was learned for each category and the positive (negative) examples were the blogs belonging (not belonging) to the category. Those blogs that do not belong to any of the top 20 categories always served as negative training/test examples. Binary weighting scheme was used in our experiment. The commonly used $tf \cdot idf$ weighting scheme yielded similar classification accuracy and we chose not to report the results due to space limitation.

[4] `http://svmlight.joachims.org/`, accessed 24 Jun 07.



(a) Precision (b) Recall

(c) F_1 (d) Train/test blog distribution

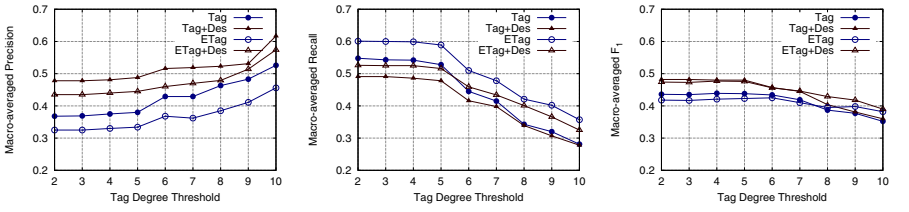
Fig. 3. Macro-averaged measures and train/test blog distribution

5.2 Experimental Results

We used *Precision*, *Recall* and F_1 to evaluate the classification accuracy. The results reported are macro-averaged measures [12].

To evaluate whether more tags lead to more accurate classification results, we obtained the classification results using different tag degree thresholds. For instance, if tag degree threshold is 5, then only those blogs having no less than 5 normalized tags will be involved in the classification. The number of train/test blogs against each tag degree threshold is given in Figure 3(d). It is clear that once the tag degree is above 5, the number of blogs in both training and test reduced sharply.

As shown in Figure 3(a), Tag together with Des, i.e., Tag+Des, achieved the best precision compared with either Tag or Des alone. With any of the three types of features, precision increased along with the tag degree threshold. Tag achieved better precision than Des when tag degree was above 5. That is, more tags led to better precision. More tags, however, led to poorer recall as shown in Figure 3(b). Recall degraded sharply when tag degree is more than 5. Among all three types of features, Tag achieved the best recall and Des was the worst.



(a) Precision (b) Recall (c) F_1

Fig. 4. Macro-averaged measures with tag expansion

As a combined measure, F_1 show that Tag+Des achieved the best classification accuracy, followed by Tag. To summarize:

- Tag was more effective than Des in blog classification despite that average number of terms in Tag is half of that in Des.
- Tag combined with Des achieved the best classification accuracy.
- More tags led to better precision but poorer recall.

To evaluate the effectiveness of tag expansion, we applied the tag expansion algorithm to expand tags of blogs involved in both training and test. We used $\alpha = 0.15$, set number of iteration to be 2, and selected those tags having score $s^2(t_i) \geq 0.15$ as expanded tags. Study of the impact of number of iterations and score threshold is out of the scope of this paper due to space limitation. Figure 4 reports the performance of expanded tag (i.e., ETag) and ETag+Des. Results obtained using Tag and Tag+Des are plotted in the Figure for easy reference. From the results, it is clear that the expanded tag led to better recall but worse precision and slightly worse $F1$ when tag degree was less than 7. It is interesting to observe that when tag degree was high (e.g., ≥ 8) tag expansion achieved better F_1 compared with the non-expanded features. This may suggest that when too many tags are given to a blog, the tags are more specific to the blog and become less effective in determining the category of a blog. Nevertheless, the observation made from this experiment requires further study to better explain the reason behind, which is part of our future work.

6 Conclusion

We studied the problem of automatically classifying tagged objects (e.g., blogs) into pre-defined categories. Compared with title and description, which are often available for many objects, tags were more effective for accurate classification. Nevertheless, our experiments suggested that more tags did not necessarily lead to better classification, which calls for further study to better explain the reason behind. We have also evaluated a tag expansion algorithm which could improve the recall but hurt precision. As tags attached to blogs may not be strongly related to any particular posts, we believe our study could benefit the research on tags attached with other online objects including pictures, video and others.

References

1. Berendt, B., Hanser, C.: Tags are not metadata, but just more content - to some people. In: Proc. of Int'l Conf. on Weblogs and Social Media (ICWSM 2007), Colorado, USA (2007)
2. Brooks, C.H., Montanez, N.: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proc. of WWW 2006, Edinburgh, Scotland, pp. 625–632 (2006)
3. Dumais, S., Chen, H.: Hierarchical classification of web content. In: Proc. of SIGIR 2000, Athens, Greece, pp. 256–263 (2000)
4. Fogaras, D., Rácz, B., Csalogány, K., Sarlós, T.: Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. Internet Mathematics 2(3), 333–358 (2005)
5. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. In: Proc. of WWW 2004, New York, pp. 491–501 (2004)
6. Hayes, C., Avesani, P., Veeramachaneni, S.: An analysis of the use of tagging in a web blog recommender system. In: Proc. of IJCAI 2007, Hyderabad, India, pp. 2772–2777 (2007)
7. Jeh, G., Widom, J.: Scaling personalized web search. In: Proc. of WWW 2003, pp. 271–279. ACM Press, New York (2003)
8. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Proc. of 10th European Conf. on Machine Learning, Chemnitz, Germany, pp. 137–142 (1998)
9. Kolari, P., Finin, T., Joshi, A.: Svms for the blogosphere: Blog identification and splog detection. In: Proc. of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (2006)
10. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: Proc. of ACM HYPERTEXT 2006, Odense, Denmark, pp. 31–40 (2006)
11. Ni, X., Xue, G.-R., Ling, X., Yu, Y., Yang, Q.: Exploring in the weblog space by detecting informative and affective articles. In: Proc. of WWW 2007, Banff, Alberta, Canada, pp. 281–290 (2007)
12. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
13. Sood, S., Owsley, S., Hammond, K., Birnbaum, L.: Tagassist: Automatic tag suggestion for blog posts. In: Proc. of Int'l Conf. on Weblogs and Social Media (ICWSM 2007), Colorado, USA (March 2007)
14. Sun, A., Lim, E.-P.: Web unit mining – finding and classifying subgraphs of web pages. In: Proc. of ACM CIKM 2003, New Orleans, LA, USA, pp. 108–115 (2003)
15. Sun, A., Lim, E.-P., Ng, W.-K.: Web classification using support vector machine. In: Proc. of 4th WIDM held in conj. with CIKM 2002, Virginia, USA (2002)