VIDEO DATABASES ANNOTATION ENHANCING USING COMMONSENSE KNOWLEDGEBASES FOR INDEXING AND RETRIEVAL

Amjad A. Altadmri
Computing Department
University of Lincoln
Lincoln, UK
atadmri@lincoln.ac.uk

Amr A. Ahmed
Computing Department
University of Lincoln
Lincoln, UK
aahmed@lincoln.ac.uk

ABSTRACT

The rapidly increasing amount of video collections, especially on the web, motivated the need for intelligent automated annotation tools for searching, rating, indexing and retrieval purposes. These videos collections contain all types of manually annotated videos. As this annotation is usually incomplete and uncertain and contains misspelling words, search using some keywords almost do retrieve only a portion of videos which actually contains the desired meaning. Hence, this annotation needs filtering, expanding and validating for better indexing and retrieval.

In this paper, we present a novel framework for video annotation enhancement, based on merging two widely known commonsense knowledgebases, namely WordNet and ConceptNet. In addition to that, a comparison between these knowledgebases in video annotation domain is presented. Experiments were performed on random wide-domain video clips, from the *vimeo.com* website. Results show that searching for a video over enhanced tags, based on our proposed framework, outperforms searching using the original tags. In addition to that, the annotation enhanced by our framework outperforms both those enhanced by WordNet and ConceptNet individually, in terms of tags enrichment ability, concept diversity and most importantly retrieval performance.

KEY WORDS

Knowledgebased Systems, Commonsense Knowledgebase, Computer Vision, Video Indexing, Video Semantic Annotation.

1 Introduction

Due to the availability of inexpensive hand held digital cameras and popularity of video sharing websites, the number of video clips uploaded everyday increases in noticeable way. As a result, the need for intelligent management tools for these data become more pressing, like efficient storage, semantic annotation, rating, indexing and retrieval.

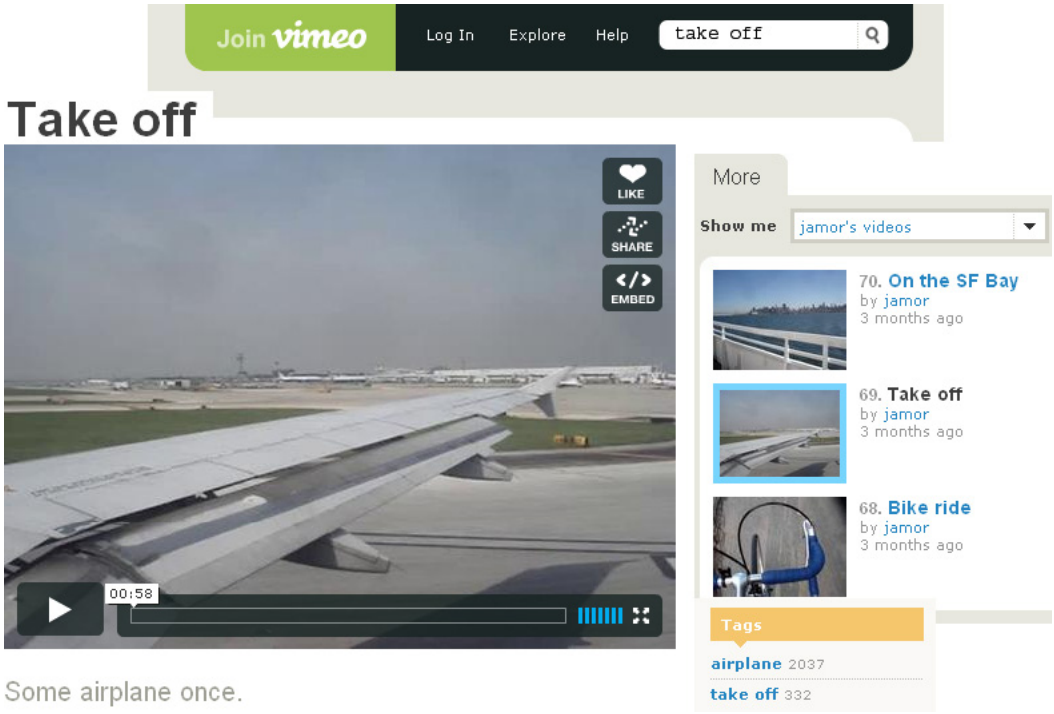
These video collections contain all types of manually annotated videos. This manual annotation is usually incomplete, uncertain and has multiple linguistic errors such as misspelling. They are also subject to different versions of the same language, such as British and American English. In addition to that, search using some keywords will mostly retrieve partial results, which are only videos that contain the exact search keywords. As a result, this annotation needs filtering, expanding and validation for better indexing and retrieval.

Work in visual annotation differs from traditional text mining application. While the description in video annotation is mainly few separated words, the analysis in text mining is usually performed on documents that contain full meaning sentences. In addition to that, almost visual annotation is about visual events/objects involved, while in text the sentences contain all possible topics.

While a number of approaches try to link between low-level features and semantic meaning, using manually annotated videos, others focus on the semantic level to link existing annotation concepts to indicate existence of other different concept in a video clip.

In this paper, a framework for video annotation validating and enriching is proposed. This framework combines two widely known commonsense knowledgebases in text mining field, namely WordNet[4] and ConceptNet[12], to enhance retrieval performance. In addition to that, a comparison between properties of WordNet and ConceptNet is presented from visual applications point of view.

Experiments were performed on random wide-domain video clips from *vimeo.com* website, which is a personal contributed unstructured video website. A snap shot of its interface is depicted in figure 1. The results show that searching for a video over enhanced tags using the proposed framework outperforms searching using the original tags. In addition to that, annotation enhanced by our framework outperforms both these enhanced by WordNet or ConceptNet individually.



take off

Take off

Show me jamor's videos

70. On the SF Bay
by jamor
3 months ago

69. Take off
by jamor
3 months ago

68. Bike ride
by jamor
3 months ago

airplane 2037

take off 332

Some airplane once.

Figure 1. A snapshot from vimeo.com website.

The rest of the paper is organized as follows: In section 2, the key related work is discussed. A comparison between WordNet and ConceptNet, in the visual contents field, is presented in section 3. Our novel framework is proposed in section 4. Then, experiments, results and evaluation are described in section 5. The paper is finally concluded in section 6, where future work is also suggested.

2 Previous work

In this section, the key related work is reviewed, where the focus is mainly on ”generalize-able” approaches in wide-domain video applications. In other words, video semantic annotation or retrieval systems that have been built without depending on special domain features, like in [6]. Bertini *et al.*[2] presented a learning approach for rules of events in video clips. The output of this approach is presented in a generalized Ontology. Other approaches tried to use association mining techniques to learn the connection between concepts to indicate existence of high-level concept from existence of others [13, 7]. Other approaches [3, 1] have directly included visual knowledge in multimedia domain-specific Ontology, in a form of low-level visual descriptors for concept instances, to perform semantic annotation.

As these methods almost depend on rules that are created by domain experts, they become not practical for manipulating large set of rules and provide less efficiency in wide domain. In addition to that, they are subject to some inconsistency, inherited from variations of the involved humans' culture, mood, personality as well as the specific topic.

Research in text mining area manages to build considerable commonsense knowledgebases. Commonsense is identified as the information and facts that are expected to be commonly known by ordinary people. WordNet [4], Cyc [11] and ConceptNet [12] are considered to be the widest commonsense knowledgebases currently in use.

In video annotation area, these knowledgebases have recently received more attention for solving annotation issues, by finding related concepts. In [15] concepts' relationships in public video databases are learned using ConceptNet's *”get context”* functionality. In addition to that, in [14], a user creates a visual concept for a group of images supported by WordNet, then ConceptNet is used to calculate the distance between the concepts to find semantically related annotations. On the other hand, some researchers in text retrieval area merge results, obtained individually, from the ConceptNet and the WordNet to achieve better query expanding [9].

In our work, WordNet and ConceptNet are combined, to utilize their strong functionalities. First, the ambiguity in words level is resolved using WordNet. Then, each video's tags are validated by exam-

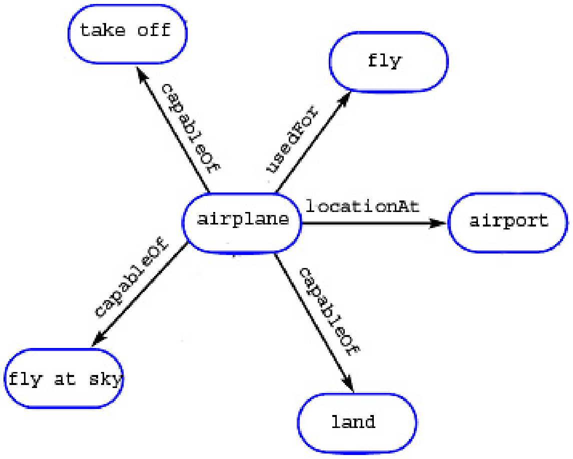


Figure 2. A snapshot of ConceptNet relationships.

ining their mutual relationships via ConceptNet, as explained in section 4.1 and 4.2.

3 ConceptNet vs. WordNet

In this section, a brief introduction to the utilized commonsense knowledgebases is presented.

3.1 ConceptNet:

ConceptNet[12] is currently considered to be the largest commonsense knowledgebase [10, 12]. It is composed from more than 700,000 free text contributors' assertions. Its nodes' core structure is concepts, which each of which is a part of a sentence that expresses a meaning. ConceptNet is a very rich knowledgebase for several aspects: First, it contains a huge number of assertions and nodes. Second, it has a wide range of information. Finally, it has various types of relationships including descriptions parameters. Figure 2 presents a snapshot that includes useful relationships in visual field.

In the last version of ConceptNet *"ConceptNet 3"* [8], each relationship has several fields expressing its score, polarity and generality. This information is automatically concluded by analyzing the frequency of the sentences that produced this relationship.

3.2 WordNet:

WordNet is a very rich non-domain-specific knowledgebase composed of lexical units, which consist of several synonym words. This knowledgebase gained wide popularity and usage due to its ease of use and wideness of trusted laboratories entered information[4]. In addition to that it has rich abstraction taxonomies. Figure 3 shows an example of a tree resulted by selecting synonym sets for the word *"fly"* and their hypernym sets.

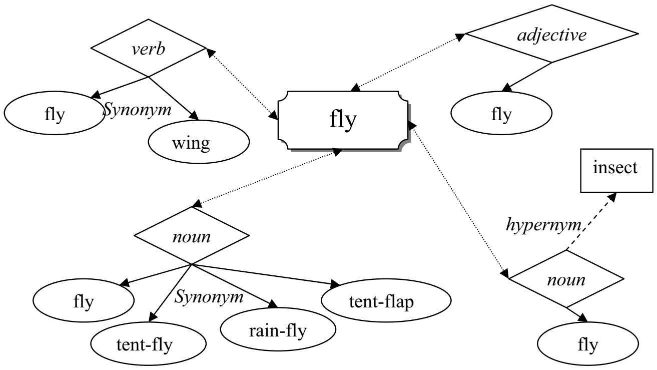


Figure 3. An example of tree built for one tag based on WordNet.

4 Proposed Framework

In this section, the proposed framework for video clips annotation enhancing for indexing and retrieval purposes is introduced. ConceptNet and WordNet are selected to be utilized in this work for several reasons:

- Both nets are general-purpose, which suits wide-domain videos.
- Both nets have natural language form, which make them easier to be compared with the annotation directly.
- Both have semantic relational structure.
- While ConceptNet nodes mainly address everyday life, WordNet focuses mainly on formal taxonomies. For example, while in ConceptNet *"dog isA pet"*, in WordNet *"dog isA mammal"*.
- While there is no connection between sentence parts in WordNet, in contrast, ConceptNet has connection between objects and their events, and objects and their locations.
- *"Synsets"* relationship in WordNet gives almost equal meaning words with little amount of abstraction, which is useful in many situations in our processing. But in contrast, ConceptNet's *"isA"* relationship is a mixture between abstraction and equality and sometimes just a property of a node. It is therefore neither symmetric relationship, to be considered as synonym, nor fully asymmetric, to be considered as abstraction.

In the following subsections, the proposed framework is described in details.

4.1 Annotation Expanding:

The aim of this stage is to find all the words that have a similar meaning to the existing tags, because they indicate to the same elements in the scene but in different words. To achieve this, first, each tag is spelling checked. Then, as in the selected site usually each tag consists of one word or a small incomplete phrase, each tag is separately expanded to its synonyms' sets using WordNet.

Each synonym set in WordNet contains similar meaning nodes. This is regardless the variety of names used for similar objects (e.g. car, automobile), the way of describing events or actions (e.g. speed up, accelerate, gain speed) and different spelling in various versions of the language (e.g. "aeroplane" in the British English, and "airplane" in the American English). In other words, each synonym set contains words in the same level of meaning.

Each tag is represented in this stage by multiple sets; each of these sets has a type, which is either a *noun*, a *verb*, an *adjective* or an *adverb*. Hence, each set represents a part of the scene either an object, action or even a description of one of them. Some words belong to more than one type of sets, which is mainly to distinguish some verbs and nouns with similar format. For example, in figure 3 The word *fly* plays many rules, one as a verb for airplane, and other as a noun which is an insect.

In addition to that, hyponym sets are generated for each synonym set. Hence in this stage, each initial annotation entry is expanded to multiple sets that assigned types, and multiple abstract sets. As a result, each tag forms a kind of tree.

These trees are very rich comparing to the initial annotations' entries for several reasons:

- As people usually annotate video clips according to their own expressions, and search also in the same way, this expansion enables searching over different spelling in different versions of the same languages. For example, *"armored"* and *"armoured"* in British and American versions of English respectively.
- Searching for a concept by different words like *"exam"* and *"test"*.
- Searching for concept using abstraction. For example, in spite of people do not tend to annotate a clip that has a *"car"* using the word *"vehicle"*, but it is highly expected that searching for *"vehicle"* should return all videos containing cars. This is achieved using our expansion tree, in contrast to the difficulty of achieving that through the initial tags.

As a result, it is clear that this expansion highly increases the number of tags, but it is noticeable that not all of these tags are valid. Hence, the next step is to validate these candidate tags.

4.2 Annotation Validation

In this layer, ConceptNet[12] relationships are used to intersect all annotations' trees, which resulted from the previous expansion step (section 4.1), to validate each branch of these trees. In addition to that, a certainty score corresponding to frequency of relevant actions in real life is given.

A full intersection operation is applied between nouns' lists and verbs' list using the ConceptNet's *"capableOf"* and *"usedFor"* relationships. This intersection selects more expressing synonym sets and deletes the other. Same operation between nouns' list is repeated using *"locationAt"* relationship.

Finally, a deeper weighting operation is performed for annotations which belong to more than one synonym set to reduce ambiguity. First, all these lists are given an equal weight and a spreading process is performed for each entry over its neighbors. An intersection operation between one from other annotations and one from these resulted trees is performed then repeated for all annotations. Any set intersects with other annotations will be marked as a represented set. The set that has the highest weight will be selected according to *"the winner takes all"* rule, and the other are deleted. New nodes' weights will be calculated according to equation 1.

$$N_w = A_w \times R_s \tag{1}$$

Where: N_w is new node weight, A_w is initial annotation weight and R_s is relation score.

For achieving the best results from intersection, the comparison process will be applied after ConceptNet nodes' analyzing, which aims to obtain each node's core.

This is done by parsing each node's words using Stanford parser [5], and deleting non-useful parts in visual field. The non-useful parts in this area vary from some prepositions and stop words to some adjectives and adverbs. For example "fast" is a useful adjective in visual area because it holds a meaning related to motion, but "better" is not as it does not reflect of low-level visual features in an agreed way. Hence, the nodes are compared more effectively.

5 Experimental results

Experiments were performed on random wide-domain video clips from the *vimeo.com* website, which is a personal contributed video website. These experiments were executed on 627 randomly selected video clips containing 6058 tags, and the results were evaluated using Retrieval degree, Enrichment ratio and Diversity.

5.1 *Retrieval degree:*

For retrieval purposes, the average number of correctly retrieved clips for a query phrase is calculated. Initially it was 1.70 video per query, but, using our framework, the average has been increased to 5.31 video per query, figure 4.

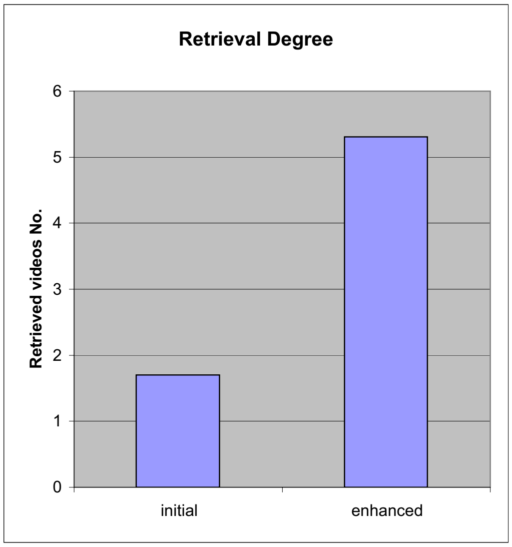


Figure 4. Retrieval degree.

5.2 *Enrichment ratio*

Tagging ratio, which is the average number of tags per video, and Enrichment ratio, which is the percentage of tagging ratio increase after enhancing annotation, formulas are explained in equations 2 and 3 respectively.

$$T = \frac{\sum_{i=1}^{N}(C_i + M_i)}{N} \quad (2)$$

Where: T is the Tagging ratio, N is the total number of videos, C_i and M_i are the number of Correct and Misspelled tags in video(i).

$$E = T2 \ / \ T1 \quad (3)$$

Where: E is the Enrichment ratio, T_1 and T_2 are the Tagging ratio before and after enhancement respectively.

As tagging ratio has risen from 9.66 tags per video clips in the dataset to 32.42 tags after annotations' enhancing, enrichment ratio has achieved a considerable degree about 336%. This is although 3.80 misspelled tags per video were removed. Figure 5 depicts the ratio of initial correct and misspelled tags to the resulted correct spelling tags.

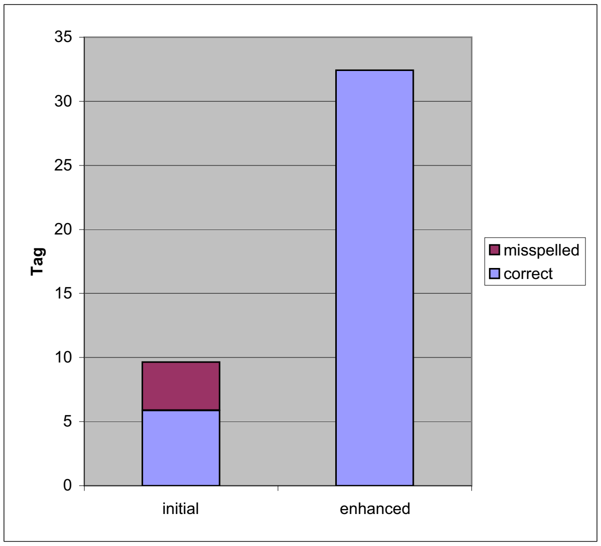


Figure 5. Enrichment ratio.

5.3 *Diversity*

The Diversity of annotations express the different topics exist in the dataset. It has been raised in a noticeable degree also from 3578 different tags in the first stage to 9271. This diversity achieves 260% increase in the topics indexed. Figure 6 demonstrate this increasing of all differentiated tags.

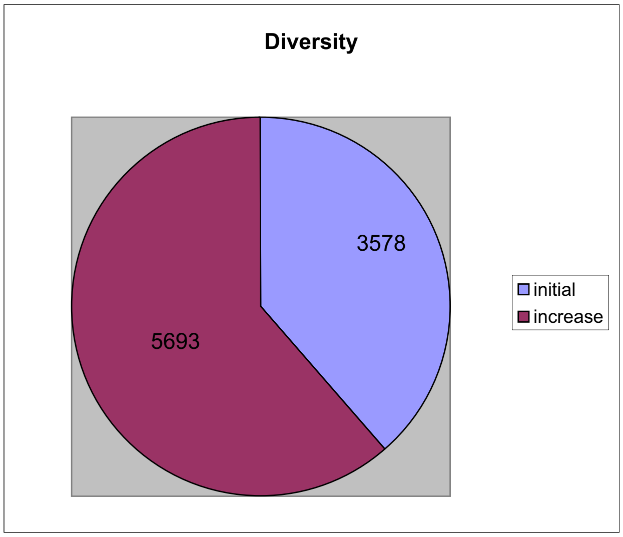


Figure 6. Diversity.

These results show that searching for a video over the enhanced tags outperforms searching using the original tags. In addition to that, annotation enhanced by the proposed framework outperforms both these enhanced by WordNet or ConceptNet individually, in terms of tags enrichment ability, Concept Diversity and most importantly retrieval performance.

6 Conclusion

In this paper, a novel automatic semantic video annotation enrichment framework is presented. This framework makes use of commonsense to enhance existing video annotation for better video indexing and retrieval. In addition to that, a comparison between used commonsense knowledgebases, WordNet and ConceptNet, is introduced from visual application area point of view. Experiments evaluation demonstrated that each one of these nets can improve the annotation of the video for a certain degree, but merging them in the proposed way outperforms each one individually. This evaluation is calculated for tags enrichment ability, Concept Diversity and most importantly retrieval performance.

First future enhancement on the proposed framework is an intelligent misspelling auto-correction. This will be performed by applying an ordinary misspelling tool to fetch all possible corrections then selecting the most suitable meaning correction for the other tags based on ConceptNet relationships.

The next main step is to explore free text analysis for annotation enhancement, which will enable the framework to be generalized to operate on any unstructured video sharing website. This will open the door towards many research directions, such as, building concepts relations models, based on learning.

Acknowledgment

The authors would like to thank Prof. Andrew Hunter for his efforts. This project is fully funded by Damascus University, Syria.

References

[1] A. D. Bagdanov, M. Bertini, A. D. Bimbo, G. Serra, and C. Torniai. Semantic annotation and retrieval of video events using multimedia ontologies. In *International Conference on Semantic Computing*, pages 713–720, 2007.

[2] M. Bertini, A. D. Bimbo, and G. Serra. Learning ontology rules for semantic video annotation. In *2nd ACM workshop on Multimedia semantics on International Multimedia Conference*, pages 1–8. ACM New York, NY, USA, 2008.

[3] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, I. Kompatsiaris, S. Staab, and M. G. Strintzis. Semantic annotation of images and videos for multimedia analysis. In *Proceedings of the 2nd European Semantic Web Conference*, pages 592–607. Springer, 2005.

[4] C. Fellbaum. *WordNet: an electronic lexical database.* Cambridge, Mass: MIT Press, 1998.

[5] S. N. Group. The stanford nlp log-linear part of speech tagger.

[6] N. Haering, R. J. Qian, and M. I. Sezan. A semantic event-detection approach and its application to detecting hunts in wildlife video. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(6):857–868, 2000.

[7] A. G. Hauptmann, M. Y. Chen, M. Christel, W. H. Lin, and J. Yang. A hybrid approach to improving semantic extraction of news video. In *International Conference on Semantic Computing, 2007. ICSC 2007.*, pages 79–86, 2007.

[8] C. Havasi, R. Speer, and J. Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, 2007.

[9] M. H. Hsu, M. F. Tsai, and H. H. Chen. Query expansion with conceptnet and wordnet: An intrinsic comparison. *Lecture Notes in Computer Science*, 4182:1–13, 2006.

[10] M. H. Hsu, M. F. Tsai, and H. H. Chen. Combining wordnet and conceptnet for automatic query expansion: a learning approach. In *Asia Information Retrieval Symposium*, volume 4993, pages 213–224. Springer, 2008.

[11] D. B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[12] H. Liu and P. Singh. Conceptnet a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.

[13] K. H. Liu, M. F. Weng, C. Y. Tseng, Y. Y. Chuang, and M. S. Chen. Association and temporal rule mining for post-filtering of semantic concept detection in video. *IEEE Transactions on Multimedia*, 10(2):240–251, 2008.

[14] B. Shevade and H. Sundaram. A visual annotation framework using common-sensical and linguistic relationships for semantic media retrieval. *LECTURE NOTES IN COMPUTER SCIENCE*, 3877:251, 2006.

[15] P. Yuan, B. Zhang, and J. Li. Semantic concept learning through massive internet video mining. In *IEEE International Conference on Data Mining Workshops*, pages 847–853, 2008.