Scientific Research Publishing

Social Networking, 2018, 7, 79-88
http://www.scirp.org/journal/sn
ISSN Online: 2169-3323
ISSN Print: 2169-3285

Framework to Classify and Analyze Social Media Content

Hayfa Aleid, Amjad Abdulrahman Alkhalaf, Aseel Helal Taemees, Hanan Saeed Almurished, Khawla Abdullah Alharbi, Nawal Hussein Alanazi

Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh, KSA
Email: haaleid@ksu.edu.sa

How to cite this paper: Aleid, H., Alkhalaf, A.A., Taemees, A.H., Almurished, H.S., Alharbi, K.A. and Alanazi, N.H. (2018) Framework to Classify and Analyze Social Media Content. *Social Networking*, **7**, 79-88.
https://doi.org/10.4236/sn.2018.72006

Received: February 25, 2018
Accepted: April 9, 2018
Published: April 12, 2018

Abstract

In the last decade, a large amount of data has been published in different fields and can be used as a data source for research and study. However, identifying a specific type of data requires processing, which involves machine learning classifying techniques. To facilitate this, we propose a general framework that can be applied to any social media content to develop an intelligent system. The framework consists of three main parts: an interface, classifier and analyzer. The analyzer uses media recognition to identify specific features. Then, the classifier uses these features and involves them in the classification process. The interface organizes the interaction between the system components. We tested the framework and developed a system to be applied to image-based social media networks (***Instagram***). The system was implemented as a mobile application (***My Interests***) that works as a recommendation and filtering system for ***Instagram*** users and reduces the time they spend on irrelevant information. It analyzes the images, categorizes them, identifies the interesting ones, and finally, reports the results. We used the **Cloud Vision API** as a tool to analyze the images and extract their features. Furthermore, we adapted ***support vector machine*** (***SVM***), a machine learning method, to classify images and to predict the preferred ones.

Keywords

Framework, Classification, Social Media Network, ***Support Vector Machine***, Machine Learning, ***My Interest***

1. Introduction

Recently, people's dependence on social media has increased dramatically. They publish and browse a huge amount of information daily. This generates a huge

data foundation to conduct scientific studies in different fields. In business, [1] shows how analyzing Twitter data can outperform usual prediction methods in marketing. However, such data must be prepared and processed before use. The data must be classified automatically using machine learning methods to filter out irrelevant content. This classification ranges from simple methods such as clustering based on word-matching to more sophisticated techniques that involve natural language processing [2].

Although social media networks are considered as a good source to conduct studies and research, their data need to be processed due to the heterogeneity of its form and source. Usually, researchers perform a sort of filtering (machine learning classification methods) to extract the appropriate data. To simplify this process, we aim to design a general framework to apply the machine learning method to social media content. This framework can be used to implement intelligent systems. For example, it can be used to classify the tweets related to a specific topic on Twitter and deduce public perceptions on a topic. We tested this framework on social media that uses images as their basic content. We developed an online mobile application (***My interest***) that serves ***Instagram***. The application aims to categorize and present the images that match the user's preferences from all the accounts she/he follows.

In the rest of the paper, we introduce an overview of the knowledge domain related to social media and machine learning. Then, we present recent work that proposes frameworks to combine machine learning with social media. After that, we provide our framework and give the development details of the ***My interest*** application.

2. Background

Recently, social media has become an integral part of our daily lives and works as a platform for exchanging experiences, opinions, and information between users via different formats such as image, text, and video. It has many advantages. For example, it can help people to enrich their knowledge and improve their social skills [3]. ***Instagram*** is a social media application used to share images and videos. Users can like the images and videos shown by other users. When a user follows other accounts, he or she can automatically see all their posts and comments. Users can also add comments and locations to images and like each other's images [4].

Image recognition is one of the oldest areas of research in the field of computing, and it provides us with a way to classify images. The ability to distinguish between a cat on a lawn and a dog on a couch and then to correctly label the animal is the main capability of image recognition. In essence, it is the process of identifying and detecting an object or a feature in an image or video [5]. **Cloud Vision API** is a free product that understands the content of images by recognizing them. The recognition is done by encapsulating powerful machine learning models using a Representational State Transfer (REST) API. This has many functions such as face detection, logo detection, and landmark detection [6].

Machine learning is one of fastest growing fields in computer science, and its objective is to give machines the ability to learn from specific input variables [7]. In this case, the definition of learning indicates performance rather than knowledge [8]. To simplify, in machine learning, "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [9]. Supervised learning is the learning of a relationship between a set of input variables and output variables, and the prediction of a successful function to form the relationship. This is used to map new input [10]. Supervised learning has two subtypes: classification and regression. In classification, the output is discrete, while regression involves real values.

A ***Support Vector Machine*** (***SVM***) is formally defined as "a supervised learning method that generates input-output mapping functions from a set of labeled training data" [12]. SVM is one of the methods of supervised learning. Developed in the 1990s [11], this method can involve classification or regression. Classification is one of the linear two-class classifiers. Its required dataset contains n labeled examples, with the example denoted by x_i for the i^{th} example in the dataset. The y_i in (x_i, y_i) for each example denotes the label associated with x_i, and x_i is the input or pattern. Using these datasets, we learn our function and improve the ability to classify the new inputs, where each label corresponds to one of the two classes. Thus, we assume that the first class is +1 (positive examples) and the second is −1 (negative examples) [12].

Therefore, let us assume we have a plane that separates examples, specifically the positive from the negative. The points x that lie on the plane satisfy

$$w \cdot x + b = 0 , \tag{1}$$

where w is normal to the plane, $|b|/\|w\|$ is the perpendicular distance from the plane to the origin, and $\|w\|$ is the Euclidean norm of w. Let $d+$ ($d-$) be the shortest distance from the separating plane to the closest positive (negative) example. The "margin" of a separating plane is defined as $d+ + d-$. For the linearly separable case, the support vector algorithm looks for the separating plane with the largest margin. See **Figure 1**. When we determine the largest margin, we obtain the best classification for the dataset [13] [14].

3. Related Work

Many studies have proposed different frameworks to apply machine learning to social media data. Most of these frameworks were proposed to overcome specific issues related to the media structure or capacity. In [15], the proposed framework focuses on how to apply machine learning on multimedia with high dimensional features and how to overcome and minimize this high dimensionality. Another work presented by Low *et al.*, provided an abstract framework that supports data mining and machine learning techniques on a large scale distributed system [16].

Some frameworks were proposed for a specific field of study. In [17], the

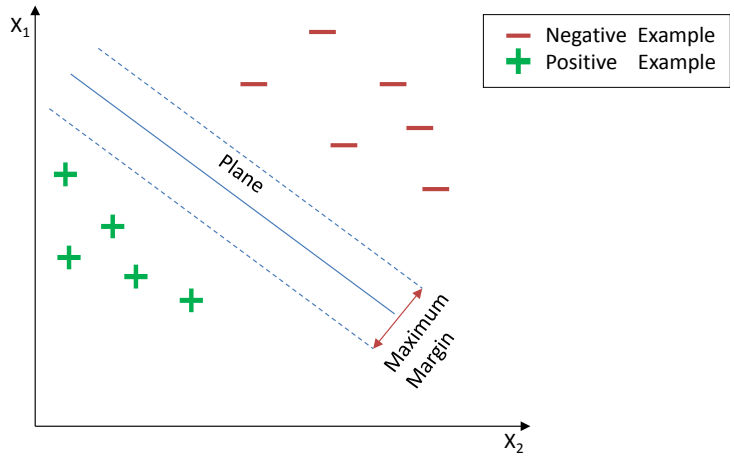

Figure 1. Support victor machine.

framework presented emphasized how to analyze and classify social media content for political consideration. Another example was provided by Geola *et al.*, which focused on applying the ***support vector machine*** to broadcast video and audio to detect if they are changed [18].

Some works, such as (Kuaa), designed a general framework that can be applied for different purposes. This automated framework was proposed and implemented based on workflow activity to execute a supervised machine learning experiment. The proposed framework consists of an interface to interact with the user, a framework core to manage the execution of the experiment and the repositories where the data is stored to perform machine learning activities.

Based on the previous literature, we aim to design an abstract framework to serve a different type of system. The framework allows these systems to adapt the machine learning method for application on social media content. Finally, we will test our framework and use it in designing and implementing a mobile web application "***My interest***".

4. Problem Statement

Recently, social networks considered as a main source of datum due to the massive amount of information that are published through them. This information is related to different kind of topics as political information and opinion, health and medical facts, market ads, social events, personal profiles, dailies etc. Therefore, many researchers rely on them to provide inputs for their studies. Unfortunately, these inputs need to be processed and classify due to their heterogeneity nature. For instance, these inputs come in several forms as texts, images or videos. Furthermore, they can be gathered from informal sources as personal blogs and accounts to more formal references as government announcements and official statistics. Thus, these data should be processed and classified to get their anticipated advantages, which increase the overhead on researchers to conduct their studies.

To make this process simple and not time and effort consuming for research-

ers, we propose a framework that facilitate the process of analyzing and classifying social media contents. The framework should be general and applicable for different type of social media. Also, it should be abstract and not restricted to any specific type of classification technique or media analyzing method. Furthermore, it should be designed into separated modules which make it expandable and can encompass extra requirement.

5. Framework Design

The framework we propose consists of three main parts: interface, analyzer, and classifier. These components work together to apply machine learning classification to social media contents. The interface role is to retrieve the media from the social network through a web call and direct them to the analyzer to obtain their features. Then, the interface sends these features to the classifier to receive the classification result. Finally, the classification results can be used for the application's main purpose. The framework is designed with a high level of abstraction as shown in Figure 2. This abstraction gives the developer the ability to add extra components or extra transactions between them if required.

The analyzer's main role is to extract specific features to be used in the classification process. These features depend on the media type and the classification method. For example, if the media is a text, then the features could be the existence of words or phrases related to a specific topic. Additionally, the features could be colors, brightness, or the existence of a specific object if the media is an image. The type of classification method determines how to present these features, single or multiple, discrete or continuous, etc. In this framework, we intend to separate the analyzer to give the developer the choice of implementing or reusing an existing component as image recognition or a text context recognition library.

The last part is the classifier which is a machine learning component that classifies the media based on the given features. The developer should decide which type of classification method to use (supervised or unsupervised classification, etc.).

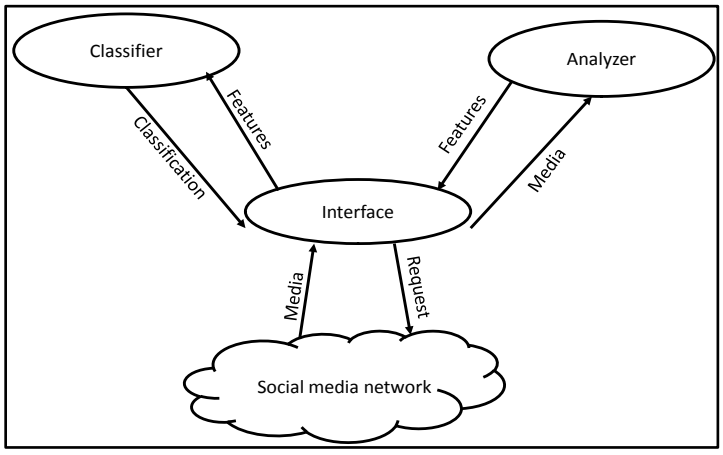

Figure 2. Framework structure.

6. *My Interests*

We tested our framework by developing a native mobile application called ***My Interests***, which works as a filtering system for ***Instagram*** accounts. The application acts as another interface for ***Instagram*** where users have the choice of browsing their accounts using the official ***Instagram*** application or our application. The main task of ***My Interests*** is to filter out images that do not match a user's preferences to save her/his time. The application infers a user's preferences by analyzing the liked images from all the accounts she/he has followed. To achieve this, we employed an image analysis and recognition tool called **Cloud Vision API**, which recognizes the contents and features of an image, such as its main color range, people's faces and emotions, as well as text, logos, etc.

Additionally, we used ***support vector machine***, a machine learning method that uses previously liked images to classify newly posted images from all followed accounts into preferred or non-preferred images. The non-preferred images are then hidden and are not shown to the user. Therefore, we assume that the user already created an account on ***Instagram*** and has followed some users. Furthermore, we focus on graphical content (images) only, where video, text, and comment are not considered.

6.1. Requirement Analysis

Table 1 summarizes all the functional requirements of the ***My Interests*** application.

Table 1. Functional requirements.

Requirement ID	Requirement Description
REQ-1	***My Interests*** allows the user to use the application when he or she logs in to his or her account on ***Instagram*** by entering their ***Instagram*** username or email or phone number and the password.
REQ-2	After a correct login process, ***My Interests*** logs in the user until he/she logs out.
REQ-3	After the user logs in, ***My Interests*** shows the **recently** followed images without any filter and categories.
REQ-4	***My Interests*** allows the user to stop or apply filtering to the images anytime.
REQ-5	If the user chooses to stop filtering, ***My Interests*** shows the user **recently followed** images without a filter.
REQ-7	***My Interests*** allows the user to like an image.
REQ-8	***My Interests*** allows the user to remove like from an image.
REQ-9	***My Interests*** allows the user to stop or classify the images into categories anytime.
REQ-10	***My Interests*** allows the user to apply filtering to the recently followed images to find the user's interests.
REQ-11	***My Interests*** allows the user to hide or show ads' images anytime.
REQ-12	***My Interests*** allows the user to show or hide the recommended places that match his/her interests.
REQ-13	***My Interests*** allows the user to log out.

To insure Security and privacy, ***My Interests*** does not view any information for a user without a valid login to his or her official account on ***Instagram***. The application interface is similar to the ***Instagram*** graphical interface to make it simple and clear for the users.

6.2. System Architecture and Design

My Interests' general architecture extends the framework we proposed. It employs the **cloud vision API** and SVM library as the analyzer and classifier. The application itself represents the interface component in the framework. Additionally, ***My Interests*** requires a local database to store and retrieve images with their features. It also connects with ***Instagram*** through web calls to log into a user account and get its contents as described in Figure 3.

6.3. System Implementation and Testing

My Interests application was developed as an object-oriented mobile application that was implemented in the Java language using Android Studio. In ***My Interests*** we used the **Cloud Vision API** tool to extract image features. We also used the SVM library to classify the images depending on the features. We developed a database to store images and their features. Our database was implemented in MySQL and placed on a web server. We tested ***My Interests*** by considering each interface page in the application as a unit. We designed a detailed test plan, and the application passed all the test cases. Then, we integrated all the components and the application passed the regression test. The following Table 2 summarizes the test cases.

We tested the performance of ***My Interests*** based on the time complexity of the system response and its reliability. The complexity was calculated using big-O notation for the main or critical transactions in the application. The

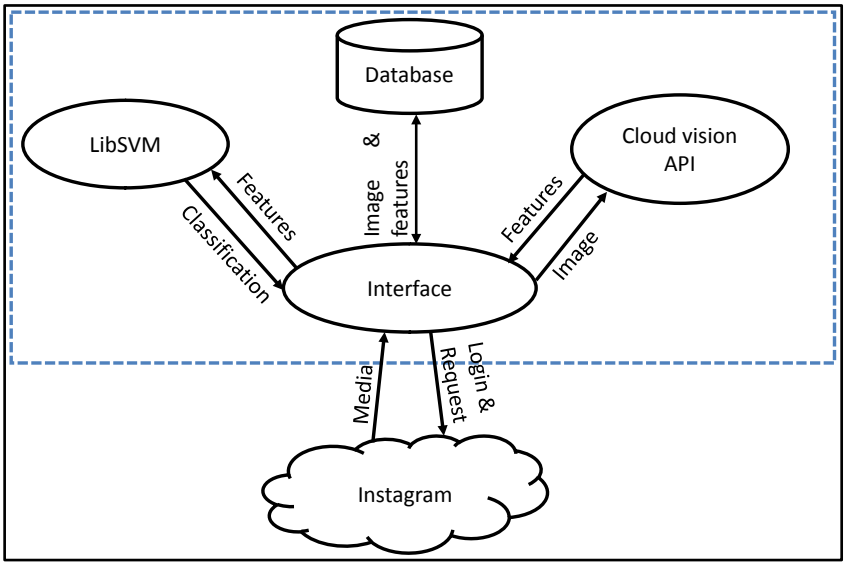

Figure 3. ***My interest*** system architecture.

Table 2. Test cases result.

Component	Testing Details	Result
Login	This component was tested using the JUnit framework	Pass
Forget Password	This component was tested manually. If the user enters invalid login information, then a message will show, and she/he can press Forget? to reset the password.	Pass
Filter Images	This component was tested manually, and it has two cases. First, when the user presses filter images, if he/she doesn't like images, then an alert will show to the user. Second, when the user presses filter images, and he/she likes the images. Then, SVM will classify the recent images and predict if the image belongs to the user's interest or not.	Pass
Stop Filtering	This component was tested manually. When the user presses again on filter images, then ***My Interests*** will show the recently followed images without a filter.	Pass
Like Image	This component was tested using the JUnit framework	Pass
Hide Ads	This component was tested manually. When the user presses hide ads, the icon will change color from red to black, and all ad images will be hidden.	Pass
Show Recommended Images	This component was tested manually. When the user presses the recommended place icon in the navigation bar, it shows all recommended images depending on the location of the liked images.	Pass
Categories Images	This component was tested manually. When the user presses the category icon in the navigation bar, it shows several images under the category.	Pass
Log Out	This component was tested manually. After the user presses the log out icon, the application redirects the user to the login UI.	Pass

complexity of extracting information from ***Instagram*** is O(n) where n represents the number of content units (images). Addition, the time complexity to classify images using linear SVM is O(m*l), where m represents the number of training examples and l is the number of features. We also measured the accuracy of ***My Interests***, and the result was satisfactory where 85% of the images were classified correctly as preferred images which are reasonable for a machine learning method.

After finishing the functionality and performance tests, we developed a brief questionnaire to test user acceptance of the application. The questionnaire was answered by a group of 45 people selected randomly from mobile device users. After receiving the feedback, the result was positive as shown the **Figure 4**.

7. Conclusion

The demand for applying machine learning classification to social media content has increased and become a necessity for many studies and surveys. We intended to make this step simple and applicable by developing a framework for developing software that applies machine learning to social media content. The framework is very abstract and simple. It is decomposed into three basic

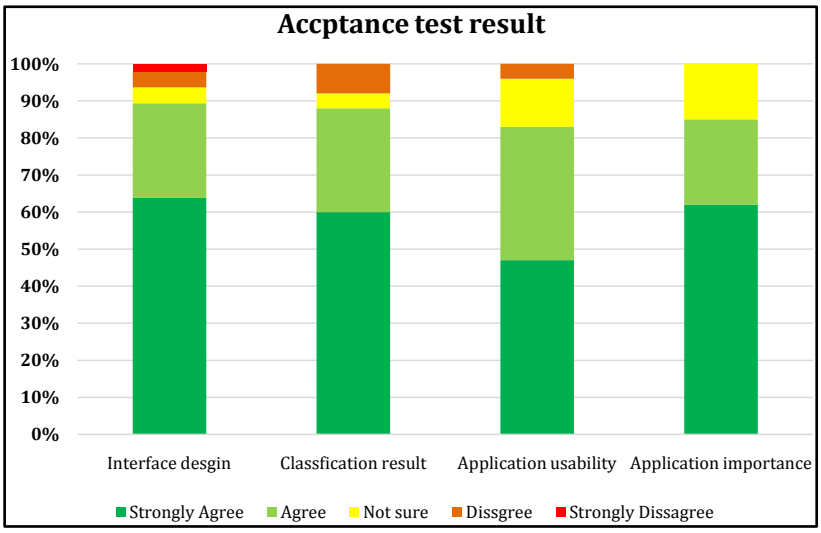

Figure 4. ***My interest*** acceptance text graph.

components: classifier, analyzer and interface. This design allows it to extend to include additional components as required. Future work will focus more on each component and give different versions that apply to several types of applications.

Based on the proposed framework, we developed an Android-based mobile application called ***My Interests***, which works as a filtering system for ***Instagram*** accounts. The application offers several services such as: filtering the images to find the user's preferences, hiding ads, categorizing images into different types and showing recommended places to the user. The application was tested by random users through simple questionnaires, and the feedback was acceptable and showed positive results. According to the questioners' feedback, we intend to consider some suggestions in the next version of this application. Currently, the system supports English language only and works on Android platforms. In the future, ***My Interests*** may support other language and may be available to other mobile system as IOS. In addition, we may apply filter not only on the Images but also on the videos and take comments on consideration. Also, ***My Interests*** can be developed to consider other popular social networks such as Twitter, Flicker, Facebook etc.

References

[1] Asur, S. and Huberman, B.A. (2010) Predicting the Future with Social Media. *Proceedings of* 2010 *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Toronto, 31 August-3 September 2010, 492-499. https://doi.org/10.1109/WI-IAT.2010.63

[2] Dai, X., Bikdash, M. and Meyer, B. (2017) From Social Media to Public Health Surveillance: Word Embedding Based Clustering Method for Twitter Classification. *Proceedings of SoutheastCon*, Charlotte, NC, 30 March-2 April 2017, 1-7.

[3] Wang, Q., Chen, W. and Liang, Y. (2011) The Effects of Social Media on College Students. Johnson & Wales University, Providence, RI.

[4] Manikonda, L., Hu, Y. and Kambhampati, S. (2014) Analyzing User Activities, De-

mographics, Social Network Structure and User-Generated Content on Instagram. arXiv preprint arXiv:1410.

[5] Image Recognition—MATLAB & Simulink. https://www.mathworks.com/discovery/image-recognition.html

[6] Vision API—Image Content Analysis/Google Cloud Platform. https://cloud.google.com/vision/

[7] Shalev-Shwartz, S. and Ben-David, S. (2014) Understanding Machine Learning. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781107298019

[8] Witten, I.H. and Ian, H. and Frank, E. (2005) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufman, San Francisco.

[9] Mitchell, T.M. and Tom, M. (1997) Machine Learning. McGraw-Hill, New York.

[10] Cord, M. and Cunningham, P. (2008) Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval. Springer Verlag, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7

[11] Wang, L. (2005) Support Vector Machines: Theory and Applications. Springer Verlag, Berlin, Heidelberg. https://doi.org/10.1007/b95439

[12] Ben-Hur, A. and Weston, J. (2010) A User's Guide to Support Vector Machines. In: Carugo, O. and Eisenhaber, F., Eds., *Data Mining Techniques for the Life Sciences*, *Methods in Molecular Biology* (*Methods and Protocols*), Vol. 609, Humana Press, 223-239.

[13] Osman, H. (2007) Novel Multiclass SVM-Based Binary Decision Tree Classifier. *Proceedings of* 2007 *IEEE International Symposium on Signal Processing and Information Technology*, Giza, 15-18 December 2007, 880-883. https://doi.org/10.1109/ISSPIT.2007.4458093

[14] Joachims, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Nédellec, C. and Rouveirol, C., Eds., *Machine Learning*: *ECML*-98, *Lecture Notes in Computer Science* (*Lecture Notes in Artificial Intelligence*), Vol. 1398, Springer, Berlin, Heidelberg.

[15] Gao, L., Song, J., Liu, X., Shao, J., Liu, J. and Shao, J. (2017) Learning in High-Dimensional Multimedia Data: The State of the Art. *Multimedia Systems*, **23**, 303-313. https://doi.org/10.1007/s00530-015-0494-1

[16] Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C. and Hellerstein, J.M. (2012) Distributed GraphLab: A Framework for Machine Learning in the Cloud. *Proceedings of the VLDB Endowment*, **5**, 716-727. https://doi.org/10.14778/2212351.2212354

[17] Stieglitz, S. and Dang-Xuan, L. (2013) Social Media and Political Communication: A Social Media Analytics Framework. Social Network Analysis and Mining, **3**, 1277-1291. https://doi.org/10.1007/s13278-012-0079-3

[18] Goela, N., Wilson, K., Feng, N., Divakaran, A. and Otsuka, I. (2007) An SVM Framework for Genre-Independent Scene Change Detection. *Proceedings of the* 2007 *IEEE International Conference on Multimedia and Expo*, *ICME* 2007, Beijing, 2-5 July 2007, 532-535. https://doi.org/10.1109/ICME.2007.4284704