Front. Comput. Sci. China 2010, 4(3): 324–333
DOI 10.1007/s11704-010-0102-7

REVIEW ARTICLE

Three challenges in data mining

Qiang YANG (✉)

Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China

Abstract In this article, I will discuss three challenges in today's data mining field. These challenges include: the transfer learning challenge, the social learning challenge and the mobile context mining challenge. I pick these three challenges because I think time is ripe for each of them to be addressed in a major way in the near future, given the current technological and societal readiness to tackle them. I also believe that each of the three challenges discussed in this article will help move the science and engineering of data mining forward, and have a great impact on society.

Keywords data mining, transfer learning, social learning, mobile computing

1 Overview

Today, the field of data mining has entered a new era where the type of data is becoming ever more complex and the scale of data ever larger. At the same time, the ability to mine useful data quickly and effectively from large quantities of data has become an issue of corporate and even national importance. In this article, I pick three areas of data mining where I believe we will see major breakthroughs in the near future, and where researchers are expected to gain new insights, and important applications are expected to be developed. These three areas are: data mining via transfer learning, mining social networks and social data, and mining contextual knowledge in mobile-user generated data.

2 *Transfer learning* challenge

2.1 Overview

Transfer learning aims to capture the knowledge from one or more auxiliary tasks in order to help discover patterns of importance in a different, but related, target tasks. In transfer learning, the training data from source or auxiliary domains, and the application or future data in a target domain, may follow different distributions or be represented with different features [1]. This problem is interesting because it violates some of the most fundamental assumptions of traditional machine learning and data mining, which require that the data are from the same space. Transfer learning is particularly useful in new data mining domains, where we do not have much labeled or annotated data to help us build a credible model, especially when budget or technical issues limit our ability to acquire new or high quality data labels. This difficult situation forces us to look elsewhere to find auxiliary data sources that are related to the target domain and that have plenty of labeled data, such that the auxiliary data can take the place of the labeled data in the target domain. *How to 'mine' useful knowledge from the auxiliary data sources, even when they appear to be different, is a critical challenge today in data mining.* We call this the *transfer learning* challenge. Below, we give an overview of development in this direction.

Recently, we have provided a survey of transfer learning [2], wherein we gave a brief overview of transfer learning that we summarize here. Informally, transfer learning can be defined using the notions of domain and task. A domain consists of a feature space in which to describe the attributes of the problem, alongside a marginal probability distribution of the data. In this

Received May 12, 2010; accepted June 4, 2010

E-mail: qyang@cse.ust.hk

definition, a domain does not involve a label space. Instead, label space is part of a task; that is, a learning task consists of a label space and a mapping function to be learned, which maps from problem features to labels. Using these concepts, a transfer learning problem can be defined as follows. Given a source domain and a source learning task as well as a target domain and a target learning task, transfer learning aims to enhance the performance of learning for the target task in the target domain using the knowledge gained from the source domain and the source task. What distinguishes transfer learning from traditional learning is that it is cross domain; either the source and target domains, the source and target tasks, or both, are different. A particular form of transfer learning is multi-task learning [1], which considers the symmetry between the learning problems among multiple tasks. In this formulation, each task can learn both from its own domain and from others' domains, and learning is conducted at the same time across domains.

Transfer learning techniques have been explored in several different application domains. Raina et al., [3] and Dai et al., [4,5] proposed to mine knowledge about text data and transfer this across domains to image data. Blitzer et al., [6] used structural-correspondence learning (SCL) for solving natural language processing tasks. An extension of SCL was proposed in [7] and a scaled version in [8] for solving sentiment classification problems. Wu and Dietterich [9] proposed to use both inadequate target domain data and plenty of low-quality source domain data for image classification problems. Arnold et al., [10] proposed to use transductive transfer learning methods to solve name entity recognition problems. In [11], a novel Bayesian multiple-instance learning algorithm is proposed, which can automatically identify the relevant feature subset and use inductive transfer for learning multiple (conceptually related) classifiers in a computer aided design (CAD) domain. In [12], Ling et al., proposed an information-theoretic approach for transfer learning, to address the cross-language classification problem from English to Chinese. [13] proposed to transfer across two domains with different feature spaces. Transfer learning has also been applied to important problems in industry. For example, [8] applied transfer learning to large scale sentiment analysis problems that arise when classifying users' feedback on products, where the user labeled data in one product area can benefit the learning of user feedback in related areas.

2.2 Applications

In this section, I will highlight a couple of interesting applications of transfer learning, showing areas where transfer learning has already made an impact in practice.

2.2.1 Transfer learning in wireless indoor location estimation

With the proliferation of the wireless technology, many mobile users' geolocation data such as GPS and Wi-Fi data become available. Being able to exploit the Wi-Fi data for localization has particular advantages over many other types of sensors, because Wi-Fi is cheap and pervasive, being available indoors as well as outdoors. If we can build successful Wi-Fi based location prediction models, we can then build many higher-level applications ranging from health care to logistics monitoring [14]. A machine-learning-based indoor localization model assumes that the learning happens in two phases. In the *offline* phase, labeled training data are used to train a location-prediction model. In the *online* phase, the localization model learned in the offline phase is used to locate a mobile device online. However, this two-phase flow is based on the assumption that the data distribution remains stationary, and this assumption may not hold in many real-world situations, for several reasons. First, the data distribution may be a function of time and space, making it expensive to collect the training data at all locations in a large building. Second, the data distribution may be a function of the client devices, making a model trained for one type of device (say Cisco Aironet 350) to be invalid when applied to another device (say Huawei E5830 3G).

To illustrate how transfer learning helps alleviate the recalibration problem, we consider transfer learning across *devices* for a two-dimensional Wi-Fi indoor localization problem. We denote the Wi-Fi signal data collected by *device* **A** as D^a and denote Wi-Fi signal data collected by device **B** as D^b. We assume D^a to be fully labeled whereas D^b to have only a few labeled examples and some unlabeled ones that can be obtained easily by quickly walking through the environment without providing labels on the way. In our empirical test, we collected a large amount of labeled data D^a from device **A,** whilst only collecting a little labeled data D^b from device **B**. Note that, although the devices may be different from each other, the learning tasks on these devices are related since they all try to learn a mapping function from their corresponding signal spaces to the *same* location space. This observation allows us to build a bridge between the devices. In fact, the learning problem is somewhat symmetric, in that knowledge gained from D^a can be beneficial to learning from D^b, and vice versa. This fact motivated us to solve the problem as a multi-task learning problem [1], in which each task can be both a source task and a target task. In particular, we extend the multi-task learning for multi-device localization problem by only requiring that the hypotheses learned from a *latent* feature space are similar [15,16]. In other words, we look for appropriate feature mappings, by which we can map different devices' data to a well-defined low-dimensional feature space. In this latent space, new devices can benefit from integrating the data collected previously by other devices to train a localization model. Our algorithm combines both *feature representation-transfer* and *parameter-transfer* for transfer learning.

2.2.2 Transfer learning in bioinformatics

In bioinformatics, many data mining techniques have been developed to solve biological problems due to the properties of wet-laboratory experiments. Experimental approaches are expensive, time-consuming and labor-intensive, so that there is a large gap between the quantity of annotated and non-annotated biological data. In these cases, traditional methods cannot build effective predictive models and thus will not achieve satisfactory prediction results. We illustrate how transfer learning can really help in solving the lack of data problem with the example of protein subcellular localization and protein–protein interaction.

Briefly, *protein subcellular localization* is crucial for genome annotation, protein function prediction and drug discovery [17]. Proteins perform their appropriate functions as, and only when, they localize in the correct subcellular compartments. Many methods have been developed and applied in an attempt to predict protein subcellular localization. Methods in Refs. [18–20] use amino acid composition to predict localization. Furthermore, scientists took account of sequence order alongside amino acid composition to overcome the missing information problem [21,22]. Supervised learning algorithms such neural networks [23], K nearest neighbor algorithm [24], support vector machine (SVM) [25] have been widely applied to solve this problem. Of these learning-based approaches, SVM is popularly adopted in bioinformatics and is shown to have better relative performance than many others.

In practice, we find that there is a large gap between the number of known and unkown protein subcellular localizations. For instance, according to the SwissProt database version 50.0 related on 30-May-2006, the number of protein sequences with localization annotations is approximately 14% of total eukaryotic protein entries [26]. This means that there are approximately 86% of eukaryotic protein entries without localization labels, which motivates us to find computational methods to predict the protein subcellular locations, both automatically and accurately.

Given the lack of data in any one biological domain, when we consider many different organisms together, we may gain much effectiveness in predictive performance. Take Cell-PLoc [27] as an example. This package lists six predictors: Euk-mPLoc, Hum-mPLoc, Plant-PLoc, Gpos-PLoc, Gneg-PLoc and Virus-PLoc, and these are specialized for predicting for eukaryotic, human, plant, Gram-positive bacteria, Gram-negative bacterial and viral proteins, respectively. We note that these classifiers are trained separately from one another, even though there is much common knowledge that can be shared among them. This observation motivates us to formulate a multi-task learning framework [1] to overcome the lack of annotated data problem. Our experimental results show that, by leveraging learning across different organisms for protein subcellular localization, the multi-task approach can indeed provide additional performance gain [28].

2.3 Summary

As mentioned above, transfer learning techniques are aimed at solving practical problems when we have insufficient annotated or labeled data in a domain of interest, and we also have insufficient resource to obtain such quality data. By looking for and making use of knowledge in related domains, we are able to transfer previously found knowledge from auxiliary or source domains to our target domains. One particularly useful scenario is when we have some noisy annotated data in a target domain, and are unsure how to filter this data. For example, we may hire some employees to label the data and annotate some search clickthrough log data in a search engine application, but due to different levels of training and other factors, the resultant labeled data may not be of high enough quality. In this case, it might be useful to transfer the filtering knowledge from related textual or clickthrough log data in order to obtain clean target data.

A major task for transfer learning in the future is to produce a solid theoretical characterization of when transfer learning is expected to work. It is found that in some situations, even when the source domain data appear related to the target domain data, the result of transfer learning is negative; that is, it is better off not to transfer any knowledge from the source domain than to transfer. The mechanism of this phenomenon should be explored further. Another challenging task is to characterize the types of knowledge that are shared among many tasks, so that the transfer learning algorithms can be developed in a more systematic way. Yet another task is to explore how to scale the transfer learning applications so that real world benefits can be quantified and easily applied not only by researchers, but also practitioners. With more and more applications of transfer learning being studied, I am sure that new lights will be shed on these questions.

3 *Social learning* challenge

Social networks and social media are changing the landscape of both society and computation [29]. Services such as Facebook, Twitter, Flickr, etc., allow millions of users to get online and share information. How do we discover the underlying dynamics of these data as they evolve? How do we make use of individually sparse data, collectively, to make accurate predictions on user linkage to other users, products, and communities? How do we mix human, crowd, and computing power to build better computational models of user behavior? These issues can be collectively referred to as *social learning*.

3.1 Collaborative filtering and link prediction

Given a large number of users and products linked in a social graph, where each user might have expressed linkage to other users and preferences on some products, the *link prediction problem* is: given the users' and products' historical and current linkage information, how likely it is that a particular user might be interested in a link to another user, or be willing to purchase a particular product, at a later time? In a graph setting, this problem has been explored in detail in many papers, such as Refs. [30,31]. One challenge is that each particular user might have only a very limited purchasing history, or a user might be linked to a small number of other users; this would make the prediction for future items extremely difficult. A key approach to solving this problem is to draw upon crowds of users with similar interests, and combine their wisdom in a weighted manner [32]. Applications of link prediction are widespread, including book and movie recommendations, friend recommendation on a social network as well as targeted advertising.

A popular technique for solving the link prediction problem is collaborative filtering (CF) [32], based on the assumption that like-minded users typically choose similar friends or products. CF can be categorized broadly as memory-based CF and model-based CF. The memory-based approach conducts certain forms of nearest neighbor search in order to predict the rating for particular user-item pair. A very common memory-based CF method is the user-based model, which estimates the unknown ratings of a target user based on the ratings by a set of neighboring users that tend to rate similarly to the target user. A crucial component of the user-based model is the user-user similarity measure for determining the set of neighbors. Popular similarity measures include Pearson's correlation coefficient (PCC) [31,33] and vector similarity [34]. An alternative form of the memory-based method is the item-based model [32,35], which compares items based on the ratings they received. When there are many fewer items than users in most applications, an item based model is more scalable.

Collaborative filtering can be modeled as making inferences and predictions on an extremely large sparse graph, and can thus be considered as an instance of link prediction. Graph algorithms can therefore be brought to bear to discover new insights in the network of humans and items, or social and information networks [36]. Inference in such a graph is not only limited to the CF problem but also other types of inference: how does information travel in such a graph [37]? Are there influential nodes that are super-spreaders in the graph [38]? What are the communities and how do they evolve [39]? Many of these issues can also be collectively addressed under the generic topic of *link prediction* in a network setting.

A main future focus will be on how to effectively solve the network sparsity problem. Typical networks of users and products, such as the Netflix movie dataset, have a sparsity level of having one link per 10000 missing links. In many other social network applications, this sparsity level can go up to one link per 100000 missing links. To alleviate the sparsity problem, different techniques have been proposed to fill in some of the unknown ratings in the matrix such as dimensionality reduction [40] and data-smoothing methods [41]. The model-based approach to CF uses observed user-item ratings to train a compact model that explains the given data [42–48]. In recent contests such as the Netflix prize competition, model based approaches showed superior performance on very large data sets.

A scenario in which data sparsity can be a problem is when we handle the new ratings matrix in a new domain. As an example, when we open a new online service, the rating matrices are often very sparse, which gives rise to the so-called *cold-start* problem. To alleviate these problems, we can exploit a transfer learning method for CF by pooling together the rating knowledge from multiple rating matrices in related but different task domains. This is the approach taken in [49], which relies on the observation that many recommendation websites for recommending similar items, e.g., movies, books, and music, are often somewhat related. For example, users who visit an online bookstore can be partitioned into similar groups as users who visit an online movie store. Similarly, books can be partitioned into groups just as movies can. In other words, items share some common properties (e.g., genre and style) and users share some of the same population-wide properties as well. Their intrinsic relationships at the user and item group levels, can be uncovered using clustering techniques, by which we can then transfer the rating knowledge from a dense to sparse rating matrix. In a series of works, we solve this problem on a small scale by discovering what is common among multiple rating matrices in related domains in order to share useful knowledge [49,50].

While completely automated link prediction and CF solutions are valuable, a new *mixed initiative* solution that exploits human computation through massive crowds is particularly promising. In this approach, each person provides some feedback to a system via user opinions or text or image labels and tags. Then, a computer program combines and filters the collective human labeling results. *Crowd sourcing*, as it is known in social computing, provides enormous commercial opportunities such as targeted advertisements. Amazon's *Mechanical Turk* [51] invites massive crowds to complete a labeling work for building classification and prediction models. Many applications can benefit from crowdsourcing, including language translation, text transcription, image labeling, market research, and conducting surveys. However, a potential problem is that the user produced data are often of varying qualities. There is much noise to be removed from this data. As a result, in the social learning area, our next challenge is: how to effectively integrate the massive, online, human input while taking care to distinguish between data of different degrees of quality and usefulness?

In the social learning area, yet another challenge is how to exploit the piecewise knowledge of a huge number of small and medium-sized recommendation systems to *collectively* make better decisions. While large online sites like Amazon and Google can easily access huge volumes of user data, the enormous number of smaller online business sites, which collectively constitute the long tail of the Web, are much more likely to have very sparse user data and have difficulty in generating accurate recommendations. Many different small sites often attract similar users and/or provide similar items, if not identical ones, which implies that data about such users and items can potentially be distributed among different systems. This idea is similar to distributed link prediction or CF, where each loosely connected subsystem can make their own judgments, while collectively a meta level system can integrate their solutions. In [52], a MapReduce [53] based implementation of the popular probabilistic latent semantic analysis model is described for collaborative filtering in online news personalization. The general idea behind these algorithms is to divide the data into small sections that can be handled at an individual computing node and coordinate a large number of computing nodes to achieve scalability. Thus, our question is: how to effectively design a distributed coalitional learning framework that enables a large number of heterogeneous systems to collectively make superior decisions? How to allow heterogeneous models to be integrated in a coherent whole? In addition, while doing all the computation, how do we effectively protect the privacy of various users and institutions, so that sensitive information is not leaked to the public nor exploited by unwanted parties?

3.2 Summary

The development of data mining has uncovered two opposing trends in computation. One trend is to link everything together in a linked network of data. Under this model, social and information networks can be studied, where a key problem is to predict the missing links given the existing, current and historical, link and node information, even when the networks are sparse. This has led to increasing research interests in social networks, link prediction and collaborative filtering. Furthermore, when we integrate the user generated labeling data into the networks, we find crowdsourcing to be particularly attractive. Extending this amalgamation of information, we find ourselves entering the arena of heterogeneous social and information networks, where transfer learning, a challenge discussed in the previous section, can again be exploited to solving the '*cold start*' problem. Another, more or less opposing trend, is to recognize that a particular graph may be too complex to handle. Thus, a divide and conquer methodology is adopted in distributed processing for link prediction. In the near future, we expect to see more interplay between these two methodologies, and research in social learning will become more scalable (including the ability to handle extremely sparse and noisy graphs), heterogeneous and distributed with increasing attention to protection of user privacy.

4 *Mobile-context learning* challenge

4.1 Overview

Learning mobile users' current and past trajectories and activities, and inferring users' intent are critical issues in the field of mobile and online services such as mobile-based Web search and targeted advertising. As mobile devices proliferate, data mining for mobile users' contextual information is gaining very high commercial and social value. In this section, we discuss the *mobile-context learning challenge*, a solution for which can help address how to provide different services with intelligent capabilities by mining the mobile users' activity and location patterns. These application services include mobile search, mobile shopping, and long-term, low-cost healthcare for the elderly and chronically sick patients, as well as other value-added services such as mobile targeted advertising.

A mobile user's context can be defined by their environment, preceding location and action sequences and intent. Knowing the context of a mobile user may allow us to infer the user's intent more accurately and provide the needed services on time. Consider the following application scenarios in a future world.

- Imagine an application where I might use a map application to help navigate myself through a large shopping mall. The map application can inform the user information center of my location, and I can query it to find various products. When I search with any product related queries on a phone based search engine, the search engine can provide me with location information as along side recommendations for activities based on my context what I did before sending the queries, the time of day, the number of people in the mall, etc. Relevant location and activity dependent results can be returned to me, such as shops nearby that might be offering a discount.
- I want to take a taxi to the airport and hope to find someone while avoiding the heavy traffic. I can send a message with the location tag. On receiving the message, someone may reply to me so that we can make can keep each other company in such case.
- I wish to monitor my own health on a long-term basis, so that when I go to a doctor I can pass my 'USB disk' to them, and they can then read my statistics to get my health profile. This makes it possible for doctors to provide me better treatment, especially if I am recovering from an illness, or for simply monitoring my health for aging illness such as heart disease.

The above are examples of what added benefits we can be gained from mobile users' context information. This information can only be inferred through a number of data sources. For example the users' past actions and location sequences, the external environmental conditions such as the weather, the time of year and time of day, as well as many users' global movement patterns such as the traffic conditions. All this information should be fused to obtain a coherent picture of the users' environment.

In this area, several challenges remain and are difficult to solve. Firstly, presently many significant locations have their corresponding location coordinates, such as GPS, marked in various kinds of digital maps, but the *functionalities* of these locations are still unknown. Providing such a system that could allow user provided tags to locations could significantly enrich information regarding the functions of locations.

Secondly, a context pattern mining system can serve as a basis for context-aware recommendations to the user. Sensors integrated on the mobile phone and other sensor devices can help detect much useful information such as the location of the user and some possible states the user is currently in, e.g., based on accelerometer information and the user's trajectory. Such information can be used to build the user's *context* and predict what the user will be looking for. For example, based on the patterns mined from the users' context information, we may determine that the user is likely to search for a restaurant around 12pm on weekdays.

An important function that can be enabled via context-mining capability is to provide services for activity recognition [14]. One important challenge here is how to obtain an ontology of the actions, so that we can infer users' actions at different levels of granularity. When monitoring a person's activities or performing activity recognition, an untouched problem by researchers is how to automatically build the model of a complicated activity based on prior or basic activities. For example, if we have sensors attached to arms or wrists which can detect *sweeping* activity and we have RFID sensors attached to vacuum cleaners which can detect *vacuuming* activity. However, how could we integrate and reason from the fact that the subject is performing activities like '*sweeping*' or '*vacuuming*' that the subject is probably doing a more complicated activity '*cleaning-the-house*'? Such a problem might benefit from mining the user sensor traces to obtain the underlying hierarchical taxonomy among activities.

Understanding and predicting mobile-users' behavior requires one to capture the sequential nature of human activities. Therefore, graphical models are employed to learn a model which describes the underlying nature of a subject's activity [54–56]. Besides, some of our application scenarios mentioned above include long-term health monitoring of people. The underlying problem of such a health monitoring involves identifying the abnormal activity patterns of subjects. Our previous research has already spanned the field of abnormal activity recognition from sensor readings [57,58].

Besides the research work mentioned above, we also aim to link activity *recognition* with *recommendation* in some of our recent papers [59,60].

4.2 Potential solutions to mobile context mining

In context learning research, several important challenges must be met.

First, in order to mine users' mobile patterns, we must gain a deeper understanding of a session-based intention space. Important questions to answer include: How do we define a mobile user's activity session? How long should the window size of the session be? What is a logical way to extract a user session from a long user trace? What might be inferred from a user session? What is the ontology of this space? Can these be automatically learned?

Second, when building a predictive model based on users' historical information, we note that a single user's data may be very sparse. Is it possible to leverage different users' data to gain a better solution for solving the data-sparsity problem? Is transfer learning applicable to solving the sparsity problem? What other domains are also related that can allow effective knowledge transfer for learning the context of a user?

Third, once we know the low-level sensor inputs and location traces, how can we perform accurate *activity recognition* for users' intent? Activity recognition aims to understand a user's activities from a minimal set of available sensors at different levels of granularity. Achieving this goal not only provides a more reliable and informative set of context features, but also allows for an effective compression of the user history data for later use. While some work has been attempted using specialized sensors in this area [53,54], large scale real-world applications of activity recognition are still lacking.

The data mining framework for the context mining of a mobile user's traces is likely to go beyond traditional machine learning. This learning process involves issues related to spatial, temporal, inter-person and inter-sensor learning [61]. Can we integrate sequential probabilistic techniques and scalable distributed learning methods that go beyond the traditional graphical model based methods? Can we integrate multi-view, multi-instance and multi-label learning with transfer learning, together with Markov model learning?

We expect a great impact on society to be made by the research on mobile users' personalized context learning in several aspects. In science, we expect knowledge to be gained in understanding people's mobility patterns and in correlating these patterns with information on public health, environmental protection and other important scientific arena. In engineering, we expect much progress to be made in terms of truly ubiquitous computing based on context inference that is not necessarily built on expensive hardware infrastructure but instead built on low cost, pervasive hardware. In business, many services can add value to existing business services, for example, mobile search, targeted mobile advertising, mobile shopping and so on. We have already seen such a trend on the iPhone and Android platform. We will see more to come.

4.3 Summary

With the advancement of hardware technology, it is increasingly possible nowadays to detect the location and activities of a user through sensors. A particularly difficult but very useful task is to detect the context of a user or users in a mobile setting, by mining complex data. Mobile data has a number of characteristics: it is noisy, sequential, incomplete, heterogeneous and highly unpredictable. To be able to successfully mine these data is a great challenge.

5 Conclusions

In this article, I have introduced three major challenges facing data mining research and application today. The first challenge, the transfer learning challenge, is driven by algorithmic and methodological concerns. The remaining two challenges, the social learning and mobile context learning challenges are application driven. The transfer learning challenge is motivated by a lack of high-quality labeled data in data mining, which is a serious problem facing any data mining practitioner today. The social learning challenge is brought forward by the fast growth of social media and social networks, where new means of computing such as crowd sourced data mining may become the norm in the not so distant future. Finally, the mobile context mining challenge is introduced by the great proliferation of mobile communications and sensor technology, which holds potential that ranges from online commerce to health monitoring for the elderly. I believe that addressing these three challenges will help move the science and engineering of data mining forward, and generate enormous impact on society.

Acknowledgements Qiang Yang's research was supported by Hong Kong CERG/China-NSFC Grant N_HKUST624/09. He thanks the students in his research group for enlightening discussions.

References

1. Caruana R. Multitask learning. Machine Learning, 1997, 28, 41–75
2. Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010 Available at http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.191
3. Raina R, Ng A Y, Koller D. Constructing informative priors using transfer learning. In: Proceedings of 23rd International Conference on Machine Learning, Carnegie Mellon, Pittsburgh, Pennsylvania. 2006, 713–720
4. Dai W, Xue G, Yang Q, Yu Y. Co-clustering based classification for out-of-domain documents. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA. 2007, 210–219
5. Dai W, Xue G, Yang Q, Yu Y. Transferring naive Bayes classifiers for text classification. In: Proceedings of the 22rd AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada. 2007, 540–545
6. Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning. In: Proceedings of the Conference on Empirical Methods in Natural Language, Sydney, Australia. 2006, 120–128
7. Blitzer J, Dredze M, Pereira F. Biographies, Bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic. 2007, 432–439
8. Pan S J , Ni X, Sun J T, Yang Q, Chen Z. Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of WWW. 2010, 751–760
9. Wu P, Dietterich T G. Improving SVM accuracy by training on auxiliary data sources. In: Proceedings of the 21st International Conference on Machine Learning, Banff, Alberta, Canada. 2004, 871–878
10. Arnold A, Nallapati R, Cohen W W. A comparative study of methods for transductive transfer learning. In: Proceedings of the 7th IEEE International Conference on Data Mining Workshops, Washington, DC, USA, IEEE Computer Society. 2007, 77–82
11. Raykar V C, Krishnapuram B, Bi J, Dundar M, Rao R B. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In: Proceedings of the 25th International Conference on Machine learning, Helsinki, Finland. 2008, 808–815
12. Ling X, Xue G R, Dai W, Jiang Y, Yang Q, Yu Y. Can Chinese web pages be classified with English data source? In: Proceedings of the 17th International Conference on World Wide Web, Beijing, China. 2008, 969–978
13. Yang Q, Chen Y, Xue G R, Dai W, Yu Y. Heterogeneous transfer learning for image clustering via the social Web. In: ACL-IJCNLP (2009). 1–9
14. Yang Q. Activity recognition: Linking low-level sensors to high-level intelligence. In: International Joint Conferences on Artificial Intelligence (IJCAI). 2009, 20–25
15. Pan S J, Shen D, Yang Q, Kwok J T. Transferring localization models across space. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence, Chicago, Illinois, USA. 2008, 1383–1388
16. Zheng V W, Pan S J, Yang Q, Pan J J. Transferring multi-device localization models using latent multi-task learning. In: Proceed-

ings of the 23rd AAAI Conference on Artificial Intelligence, Chicago, Illinois, USA. 2008, 1427–1432

17. Su E C Y, Chiu H S, Lo A, Hwang J K, Sung T Y, Hsu W L. Protein subcellular localization prediction based on compartment-specific feature and structure conservation. BMC Bioinformatics, 2007, 8 (1): 330–341
18. Muskal S M, Kim S H. Predicting protein secondary structure content. A tandem neural network approach. Journal of Molecular Biology, 1992, 225(3): 713–727
19. Zhou G P. An intriguing controversy over protein structural class prediction. Journal of Protein Chemistry, 1998, 17(8): 729–738
20. Zhou G P, Assa-Munt N. Some insights into protein structural class prediction. Proteins, 2001, 44(1): 57–59
21. Chou K C. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins, 2001, 43(3): 246–255
22. Liu W, Chou K C. Prediction of protein secondary structure content. Protein Engineering, 1999, 12(12): 1041–1050
23. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Research, 1998, 26(9): 2230–2236
24. Huang Y, Li Y. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics, 2004, 20(1): 21–28
25. Yu C S, Lin C J, Hwang J K. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein science : A Publication of the Protein Society, Protein Sci., 2004, 13(5): 1402–1406
26. Shen H B, Yang J, Chou K C. Euk-PLoc: An ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids, 2007, 33(1): 57–67
27. Chou K C, Shen H B. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols, 2008, 3(2): 153–162
28. Xu Q, Pan S J, Xue H H, Yang Q. Multitask learning for protein subcellular location prediction. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2010
29. Wang F-Y, Carley K M, Zeng D, Mao W. Social computing: From social informatics to social intelligence. In: IEEE Intelligent Systems, March/April. 2007, 79–83
30. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. JASIST, 2007, 58(7): 1019–1031
31. Liben-Nowell D, Kleinberg J M. The link prediction problem for social networks. In: ACM Conference on Information and Knowledge Management. 2003, 556–559
32. Breese J, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th conference on Uncertainty in Artificial Intelligence. 1998, 43–52
33. Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J. GroupLens: An open architecture for Collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. 1994, 175–186
34. Herlocker J, Konstan J A, Riedl J. An empirical analysis of design choices in neighborhood-based collaborative Filtering algorithms. Information Retrieval, 2002, 5(4): 287–310
35. Sarwar B, Karypis G, Konstan J, Reidl J. Item-based collaborative filtering recommendation algorithms. In: WWW. 2001, 285–295
36. Han J, Sun Y, Yan Y, Yu P S. Mining knowledge from databases: An information network analysis approach. In: SIGMOD Conference. 2010, 1251–1252
37. Gruhl D, Guha R V, Liben-Nowell D, Tomkins A. Information diffusion through blogspace. In: WWW. 2004, 491–501
38. Tang J, Sun J, Wang C, Yang Z. Social influence analysis in large-scale networks. In: ACM KDD. 2009, 807–816
39. Leskovec J, Backstrom L, Kumar R, Tomkins A. Microscopic evolution of social networks. In: ACM KDD. 2008, 462–470
40. Linden G, Smith B, York J. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 2003, 7 (1): 76–80
41. Goldberg K, Roeder T, Gupta D, Perkins C. Eigentaste: A constant time collaborative filtering algorithm. Information Rretrieval, 2001, 4(2): 133–151
42. Ma H, King I, Lyu M. Effective missing data prediction for collaborative filtering. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2007, 39–46
43. Rennie J, Srebro N. Fast maximum margin matrix factorization for collaborative prediction. In: Proceedings of the 22nd International Conference on Machine Learning. 2005, 713–719
44. Paterek A. Improving regularized singular value decomposition for collaborative filtering. In: Proceedings of KDD Cup and Workshop. 2007
45. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. IEEE Computer, 2009, 42(8): 30–37
46. Hofmann T. Latent semantic models for collaborative filtering. ACM Transactions on Information Systems, 2004, 22(1): 89–115
47. Jin R, Si L, Zhai C, Callan J. Collaborative filtering with decoupled models for preferences and ratings. In: ACM Conference on Information and Knowledge Management. 2003, 309–316
48. Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering. In: Proceedings of the 24th International Conference on Machine Learning. 2007, 791–798
49. Li B, Yang Q, Xue X. Transfer learning for collaborative filtering via a rating-matrix generative model. In: ICML. 2009, 617–624
50. Pan W, Xiang E W, Liu N, Yang Q. Transfer learning in collaborative filtering for sparsity reduction. In: Proceedings of the 24rd AAAI Conference on Artificial Intelligence. 2010. To appear
51. Kittur A, Chi E H, Suh B. Crowdsourcing user studies with

Mechanical Turk. In: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems (2008). CHI ’08. ACM, New York, NY, 2008, 453–456

52. Das A S, Datar M, Garg A, Rajaram S. Google news personalization: scalable online collaborative filtering. In: Proceedings of WWW. 2007, 271–280
53. Dean J, Ghemawat S. Mapreduce. Communications of the ACM, 2008, 51(1): 107–113
54. Yin J, Chai X, Yang Q. High-level goal recognition in a wireless LAN. In: Proceedings of the 19th AAAI Conference on Artificial Intelligence, San Jose, California, USA. 2004, 578–584
55. Chai X, Yang Q. Multiple-goal recognition from low-level signals. In: Proceedings of the 20 AAAI Conference on Artificial Intelligence, San Jose, California, USA. 2005, 3–8
56. Hu D H, Yang Q. Cigar: Concurrent and interleaving goal and activity recognition. In: Proceedings of the 23 AAAI Conference on Artificial Intelligence, San Jose, California, USA. 2008, 1715–1720
57. Yin J, Yang Q, Pan J J. Sensor-based abnormal human-activity detection. IEEE Trans. on Knowl. and Data Eng., 2008, 20(8): 1082–1090
58. Hu D H, Zhang X X, Yin J, Zheng V W, Yang Q. Abnormal activity recognition based on HDP-HMM models. In: International Joint Conferences on Artificial Intelligence (IJCAI). 2009, 1715–1720
59. Zheng V W, Zheng Y, Xie X, Yang Q. Collaborative location and activity recommendations with gps history data. In: WWW. 2010, 1029–1038
60. Zheng V W, Cao B, Zheng Y, Xie X, Yang Q. Collaborative filtering meets mobile recommendation: A user-centered approach. In: Proceedings of the 24rd AAAI Conference on Artificial Intelligence. 2010. To appear
61. Eagle N. Mobile Phones as Social Sensors. The Handbook of Emergent Technologies in Social Research. Oxford University Press, 2010