Cross Domain Similarity Mining: Research Issues and Potential Applications Including Supporting Research by Analogy*

Guozhu Dong
Department of Computer Science and Engineering
and Knoesis Center of Excellence
Wright State University
Dayton, Ohio 45435, USA
guozhu.dong@wright.edu

ABSTRACT

This paper defines the *cross domain similarity mining* (CDSM) problem, and motivates CDSM with several potential applications. CDSM has big potential in (1) supporting understanding transfer and (2) supporting research by analogy, since similarity is vital to understanding/meaning and to identifying analogy, and since analogy is a fundamental approach frequently used in hypothesis generation and in research. CDSM also has big potential in (3) advancing learning transfer since cross domain similarities can shed light on how to best adapt classifiers/clusterings across given domains and how to avoid negative transfer. CDSM can also be useful for (4) solving the schema/ontology matching problem. Moreover, this paper gives a list of potential research questions for CDSM, and compares CDSM with related studies. One purpose of this paper is to introduce the CDSM problem to the wide KDD community in order to quickly realize the full potential of CDSM.

1. THE CROSS DOMAIN SIMILARITY MINING (CDSM) PROBLEM

After giving a general definition of CDSM, this section discusses (a) some similarity-revealing knowledge structures that can be mined by CDSM, (b) example datasets that can be considered as input to CDSM, and (3) some data preparation considerations for CDSM.

Definition: The problem of *cross domain similarity mining* (CDSM) is, given two[1] datasets collected from two application domains, mine high-quality knowledge structures that capture structural level similarity between the two domains.

The two datasets may or may not have class labels.

Example kinds of knowledge structures that can capture structural level similarities shared by two given datasets include shared decision trees, shared Bayesian models, shared (hidden) Markov models, shared (linear) regression models, shared clusterings, shared conceptual clusterings with succinct cluster descriptions, shared rankings of objects/concepts, shared attributes/schema fragments, etc. They also include the alignable-difference type of knowledge structures for two given domains, which contain substantial similarity parts shared by the two datasets and some difference parts unique for one of the two datasets. They can also be domain-specific knowledge structures, such as shared gene interaction networks and shared biological pathways for two diseases/organisms.

Example datasets that can be used as input to CDSM include two microarray datasets, perhaps with one for a well understood cancer and another for a poorly understood cancer, or two sets of internet browsing histories for users from two countries, or two sets of DNA sequences for a binding site of interest for two bacteria, or two datasets for which one wishes to discover the kinds/properties of similarity patterns they share before selecting/applying learning transfer approaches, or two heterogeneous databases that one wishes to integrate/combine, and so on. The two datasets can be heterogeneous or homogeneous with respect to their attributes and classes.

Whether **data preparation**, and what kinds of data preparation, should be performed on the datasets used in CDSM is a decision that should be made in consultation with the domain experts depending on their goals. Data preparation may be needed if there are known cross domain differences in laboratory conditions and/or data collection technologies. Sometimes one may want to use various methods to provide likely, or even purely *hypothetical*, equivalence relationship between the classes, or between the attributes, of the two datasets, in order to discover similarity revealing knowledge structures based on the hypothetical equivalences in "what-if" studies. Data mining methods may be needed to help discover various kinds of such equivalence relationships.

2. MOTIVATIONS FOR AND POTENTIAL APPLICATIONS OF CDSM

CDSM has potential to be highly useful, since similarity-revealing knowledge structures can help support understanding transfer between a better understood domain and a poorly understood domain, support research by analogy to

*The work was supported in part by NSF IIS-1044634. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

[1]One can also consider CDSM for three or more datasets, which is omitted here to simplify the presentation.

assist scientists to deal with a novel and complex research challenge and to form potential novel hypothesis, advance the field of learning transfer, and provide assistance in schema and ontology matching and integration. We discuss each of those applications in more detail here; other motivations and applications exist but are not discussed here.

(1) **Assisting users to transfer their understanding between domains**. Consider this hypothetical scenario: Having studied a disease W extensively, John has expert knowledge on W, including how the key genes interact in W. Recently he started studying a new, poorly known, disease P. He examined some shared gene interaction relationships extracted from the two microarray gene expression datasets, D_W and D_P, for the two diseases. A particular shared relationship among three genes, g_3, g_{10} and g_{18}, got his attention. John's experience on W tells him that the three genes play an important role in the development of W, occurring in a biological pathway important for W. Since that relationship also occurs in P, John felt that those three genes may also be very important for P. He focused his effort on understanding the three genes for P and got rewarded. The shared relationship transferred John's understanding of W to his understanding of P, helped improve his understanding of P and helped him form a new hypothesis on P.

(2) **Importance of analogy, and of similar structures for analogical reasoning/creative thinking**.

- Psychology and cognitive science studies indicate that analogy plays a vital role in human thinking and reasoning[2] [3], including *creative thinking*. Kepler's discovery/exposition of the *concept of gravity*[4] was aided by analogy between gravity and light [15]; research on, and development of protection against, computer virus has been assisted by analogy to biological virus; many useful algorithms and computing concepts, e.g. hill climbing and simulated annealing, are described (and perhaps invented) with the assistance of analogy.
- Psychology/cognitive science studies also show that structural similarity is the foundation of analogy based thinking [13; 15; 12; 14; 4]. Gentner and Markman [15] "suggest that both similarity and analogy involve a process of structural alignment and mapping." Christie and Gentner [4] suggest, based on psychological experiments, that "structural alignment processes are crucial in developing new relational abstractions" and *in forming new hypothesis*.

Observe that the result of a structural alignment process is typically a shared knowledge structure between two given domains. The main aim of CDSM is to mine shared knowledge structures efficiently, to help researchers find structural alignments automatically or semi-automatically.

(3) **Helping avoid negative transfer in learning transfer**. In learning transfer [23], it is desirable to use available structure/knowledge of an auxiliary application domain to help build better classifiers/clusters for a target domain. However, sometimes when two given application domains share little structure in common, negative transfer can occur, giving worse results when the auxiliary domain is used. One can use the mined shared knowledge structures to determine the amount and type of similarity between two given applications, and decide whether to use learning transfer.

The usefulness of knowledge transfer between applications has been widely recognized in many application domains (including education, learning, cognitive sciences, biological sciences, business and economic development) and in the learning transfer area [23] of data mining/machine learning.

(4) The similarity-type knowledge mined by CDSM can also be useful for **solving the schema and ontology matching problem**, which is an important issue for semantic web, data integration, data warehousing, etc. More details on schema and ontology matching is given in Section 4.

3. RESEARCH ISSUES FOR CDSM

There are many interesting research problems for CDSM. We discuss several below.

1. Mining various types of shared knowledge structures from different types of data. As discussed earlier, example kinds of shared knowledge structures include shared decision trees, shared Bayesian models, shared (hidden) Markov models, shared (linear) regression models, shared clusterings, shared conceptual clusterings with succinct cluster descriptions, alignable differences, and domain specific shared knowledge structures such as shared gene interaction networks. The input datasets can be vector of numerical attribute values (such as microarray data), text data, sequences, graphs, time series, etc.
2. Mining small diversified set of shared knowledge structures, instead of just one shared knowledge structure. This is desirable since different shared knowledge structures can exist in a given pair of datasets, and we want to increase the chance of having some mined shared knowledge structures succeed in triggering analogy based thinking in the users, without imposing a huge cognitive processing overhead on the users. Mining a small number of shared knowledge structures that are highly different from each other can be a good approach to achieve those goals.
3. Using shared knowledge structures to enhance transfer learning. This can be achieved by using the mined shared knowledge structures to determine the way and the degree two given applications are similar to each other. If we know that we can get high quality shared decision trees between two given domains but we cannot get high quality shared naive Bayes classifiers, that information can help us select more appropriate classification models in learning transfer. If we cannot discover high quality shared knowledge structures, then the two given domains may share very little in common; that information can be used to help make the

[2]Fauconnier [12] states: "*Our conceptual networks are intricately structured by analogical and metaphorical mappings, which play a key role in the synchronic construction of meaning and in its diachronic evolution. Parts of such mappings are so entrenched in everyday thought and language that we do not consciously notice them; other parts strike us as novel and creative. The term* metaphor *is often applied to the latter, highlighting the ... poetic aspects of the phenomenon.*"
[3]Gentner and Colhoun [14]: "*Much of humankind's remarkable mental aptitude can be attributed to analogical ability.*"
[4]According to Gentner [15], Kepler, a great discoverer, was a **prolific analogizer**.

decision not to use learning transfer (to avoid negative learning transfer).

4. Using shared knowledge structures to support research by analogy. For example, we want to study how to use/select shared knowledge structures to assist scientists in research by analogy, and how to mine shared knowledge structures based on the known background knowledge of the scientists in order to better support research by analogy. We may also want to study how to select potential equivalences, including hypothetical ones, between the classes and between the attributes of two given domains. For example, we can discover similarity between attributes (genes) based on the similarity of their behavior with respect to the classes of two given diseases. Hypothetical equivalences can help scientists discover analogies in "what-if" studies.

5. Studying ways to evaluate the usefulness and the quality of shared knowledge structures. For example, for mining shared decision trees for two given datasets, useful quality factors can include the accuracy of a shared decision tree in the two datasets, the similarity between the distributions of the matching data of the two datasets at the tree nodes, the simplicity of the shared decision tree, etc.

6. Extending the study to the Cross Domain Data Mining (CDDM) area, which is more general than CDSM. The research in this area will not only mine similarity revealing knowledge structures shared by two given domains/datasets, but also mine difference revealing knowledge structures which are unique to one of the two domains. Having high quality shared knowledge structures, high quality difference-revealing knowledge structures, and alignable differences, can help provide a better picture regarding the relationship between two application domains.

4. RELATED WORK

Studies on similarity measures: Due to the importance of similarity in many data-centric tasks, much has been done on studies on similarity measures. Such studies can be divided into three groups:

- Studies on similarity measures between pairs of individual objects (e.g., [16] on time series similarity, [3] on categorical data similarity, [20] on similarity between short text segments, and [2] on the limitation of distance/similarity between objects in high dimensional space).

- Studies on similarity measures between pairs of attributes (e.g. [5]).

- Studies on similarity measures between pairs of datasets (e.g. [24; 32]).

We note that CDSM aims to mine similarity revealing knowledge structures in order to support understanding transfer, research by analogy, and learning transfer; it is not concerned with measuring similarity between objects, and it is not limited to measuring similarity between attributes or similarity between datasets.

Studies on schema and ontology matching: Schema and ontology matching [27; 28; 18] is a fundamental problem in many application domains, such as semantic web, schema and ontology integration, E-business, and data warehousing. Typically, schema/ontology matching takes as input two schemas/ontologies, each consisting of a set of discrete entities (such as tables, XML elements, classes, properties, rules, predicates), and determines various relationships (such as equivalence, subsumption) that hold between attributes or other fragments of the given entities. Sometimes data instances of the two schemas/ontologies [18] are also used as input for corpus-based schema matching, to help derive useful properties such as length and type of the domain of an attribute.
In some sense, CDSM can be viewed as a generalization of schema/ontology matching, since CDSM aims to mine similarity revealing knowledge structure beyond equivalence and subsumption relationships. It should be noted that the results of CDSM can be utilized in solving the schema/ontology matching problem.
Studies on learning transfer are mainly concerned with using knowledge extracted from auxiliary datasets to help build *better* knowledge structures in a target dataset/application. For example, one uses an auxiliary dataset to help build a classifier for a target dataset, which is more accurate than classifiers built without using auxiliary datasets. Reference [23] surveyed studies on learning transfer for classification, regression, and clustering; it categorized transfer learning under three main subsettings, namely inductive transfer learning, transductive transfer learning, and unsupervised transfer learning. Importantly, the focus of learning transfer is to build better classifiers/clusterings on given target datasets. Unlike CDSM, learning transfer is not focused on mining (shared) knowledge structures, it is not aimed at helping human users to transfer understanding between application domains, and it is not aimed at assisting human users to better perform cross-domain analogy based reasoning and to better perform research by analogy. We note that CDSM has big potential in advancing learning transfer since cross domain similarities can shed light on how to best adapt classifiers/clusterings across given domains and to avoid negative transfer. The next paragraph discusses some other studies on learning transfer that are more related to CDSM.

Studies on knowledge structure transfer: Reference [6] considered structure transfer (also called "deep transfer") for situations where source and target data are (i) from different domains/applications and (ii) described by different predicates. Reference [31] considered mining rules for cross-domain transfer. Both used cross-domain predicate/attribute mappings to capture "equivalence" between predicates/attributes; however, they did not focus on mining shared knowledge structures.

Studies on shared decision tree mining: Reference [9] motivated and studied the problem of mining shared decision trees across multiple datasets/applications. Two shared decision tree mining problems were studied, namely (a) mining shared decision trees with high shared accuracy (which is defined to be the minimum of the two accuracies of each given shared decision tree in the two datasets), and (b) mining shared decision trees with high shared accuracy and high data distribution similarity (which is based on the distribution of the classes at the tree nodes). It was argued that

mining results for the second problem are more useful. Algorithms were developed to solve both problems. Experimental results on fifteen pairs of medical microarray datasets were reported to evaluate the algorithms, together with the mined shared decision trees. Future research questions on mining shared decision trees and other shared knowledge structures were discussed. Currently the authors of [9] are working on mining small diversified set of shared decision trees and some other related research questions.

Examples of knowledge transfer have been discussed in many papers, including [17][25] concerning learning by analogy, [21][26][30] concerning task/procedure transfer, [7] concerning economic policy transfer, [29][22][19] concerning cross-species biological knowledge transfer, etc. This confirms the importance of analogy and also shows that the importance of analogy is widely accepted. These papers are mostly about using knowledge transfer, but not about mining shared knowledge structures to assist knowledge transfer.

Using structural mapping to find analogy: Reference [11] discussed how to find knowledge level analogy using textual statements as input, but it did not use observation data as input and it did not consider conducting data mining to find analogy.

Contrast data mining: The CDSM direction is related to contrast data mining and applications [10; 8]. Contrast data mining is concerned with mining patterns and models that contrast multiple classes/datasets/conditions; contrast patterns include emerging patterns [10] and contrast sets [1]. Loosely speaking, the studies on contrast mining and on shared knowledge structure mining all fall into the common theme of "comparative mining of multiple classes/datasets."

Acknowledgement: The author wishes to thank Bart Goethals (and other anonymous reviewers) for very useful suggestions that helped improve this position paper.

5. REFERENCES

[1] S. D. Bay and M. J. Pazzani. Detecting change in categorical data: Mining contrast sets. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 302–306, 1999.

[2] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *International Conference on Database Theory (ICDT)*, pages 217–235, 1999.

[3] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *SIAM International Conference on Data Mining (SDM)*, pages 243–254. SIAM, 2008.

[4] S. Christie and D. Gentner. Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, 11(3):356–373, 2010.

[5] G. Das, H. Mannila, and P. Ronkainen. Similarity of attributes by external probes. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 23–29, 1998.

[6] J. Davis and P. Domingos. Deep transfer via second-order markov logic. In *ICML*, 2009.

[7] D. P. Dolowitz and D. Marsh. Learning from abroad: the role of policy transfer in contemporary policy-making. *An International Journal of Policy and Administration*, 13:5–23, 2002.

[8] G. Dong and J. Bailey, editors. *Contrast Data Mining: Concepts, Algorithms, and Applications*. Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC, to appear in 2012.

[9] G. Dong and Q. Han. Mining shared decision trees across multiple datasets. Technical Report, Department of CSE, Wright State University, 2011.

[10] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 43–52, 1999.

[11] B. Falkenhainer, K. D. Forbus, and D. Gentner. The structure-mapping engine: Algorithm and examples. *Artif. Intell.*, 41(1):1–63, 1989.

[12] G. Fauconnier. *Mappings in Thought and Language*. Cambridge University Press, 1997.

[13] D. Gentner. Structure mapping: A theoretical framework for analogy. *Cognitive Science*, 7:155–170, 1983.

[14] D. Gentner and J. Colhoun. Analogical processes in human thinking and learning. In B. Glatzeder, V. Goel, & A. von Mller (Vol. Eds.)*, On Thinking: Vol. 2. Towards a Theory of Thinking. Springer-Verlag*. 2010.

[15] D. Gentner and A. B. Markman. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56, 1997.

[16] E. J. Keogh, K. Chakrabarti, M. J. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowl. Inf. Syst.*, 3(3):263–286, 2001.

[17] M. Klenk and K. Forbus. Domain transfer via cross-domain analogy. *Cognitive Systems Research*, pages 240–250, 2009.

[18] J. Madhavan, P. A. Bernstein, A. Doan, and A. Y. Halevy. Corpus-based schema matching. In *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, pages 57–68, 2005.

[19] L. A. McCue, W. Thompson, C. S. Carmack1, and C. E. Lawrence. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Research*, 2002.

[20] D. Metzler, S. T. Dumais, and C. Meek. Similarity measures for short segments of text. In *Advances in Information Retrieval – 29th European Conference on IR Research (ECIR)*, pages 16–27, 2007.

[21] L. R. Novick. Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14:510–520, 1988.

[22] V. Olman, H. Peng, Z. Su, and Y. Xu. Mapping of microbial pathways through constrained mapping of orthologous genes. In *IEEE Computer Society Computational Systems Bioinformatics Conference*, pages 363–370. IEEE Computer Society, 2004.

[23] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10), 2010.

[24] S. Parthasarathy and M. Ogihara. Exploiting dataset similarity for distributed mining. In J. D. P. Rolim, editor, *Parallel and Distributed Processing (IPDPS) Workshops*, volume 1800 of *Lecture Notes in Computer Science*, pages 399–406. Springer, 2000.

[25] D. N. Perkins and G. Salomon. Transfer of learning. *International Encyclopedia of Education*, 1992.

[26] P. Pirolli and M. Recker. Learning strategies and transfer in the domain of programming. *Cognition and Instruction*, 1994.

[27] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.

[28] P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. pages 146–171, 2005.

[29] Z. Su, P. Dam, X. Chen, V. Olman, T. Jiang, B. Palenik, and Y. Xu. Computational inference of regulatory pathways in microbes. In *IEEE Computer Society Bioinformatics Conference*, pages 631–633. IEEE Computer Society, 2003.

[30] M. E. Taylor and P. Stone. Behavior transfer for value-function-based reinforcement learning. *Proc. International Joint Conference on Autonomous Agents and Multiagent Systems*, 2005.

[31] M. E. Taylor and P. Stone. Cross-domain transfer for reinforcement learning. In *ICML*, pages 879–886, 2007.

[32] J. Vreeken, M. van Leeuwen, and A. Siebes. Characterising the difference. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 765–774, 2007.