

Báo cáo nội dung Tuần 1

Phan Công Minh <19021330>

Mỗi sinh viên lựa chọn 10 thuật ngữ mà bạn cho là khái niệm quan trọng trong khai phá dữ liệu, bao gồm tiếng Anh, tiếng Việt và giải thích. Gửi lại cho GV phụ trách nhóm mình.

Data: dữ liệu

Data := a meaningless point in space and time, without reference to either space or time, without meaningful relations to anything else. An event out of context, a letter out of context, a word out of context. A key concept here being "out of context". ¹

When there is no context, there is little or no meaning. So, we create context but, more often than not, that context is somewhat akin to conjecture, yet it fabricates meaning. ¹

Dữ liệu: một điểm không có ý nghĩa trong không gian và thời gian, không tham chiếu tới thời gian hoặc không gian, hay nói đơn giản là không có ngữ cảnh ("context")

Thiếu ngữ cảnh, dữ liệu không có hoặc có ít ý nghĩa. Do đó, chúng ta phải "tạo" ngữ cảnh, và nhiều khi "ngữ cảnh" đó chỉ dựa trên sự phỏng đoán, song nó vẫn tạo nên ý nghĩa.

Information: thông tin

Information:

- is an understanding of the relationships between pieces of data, or between pieces of data and other information,
- generally does not provide the why (the data is what it is), nor the how (the data is likely to change over time),
- is relatively static in time and linear in nature,
- has a great dependence on context for its meaning and with little implication for the future. ¹

Thông tin:

- sự hiểu quan hệ của các mẫu dữ liệu với nhau, hoặc giữa các mẫu dữ liệu và các thông tin khác,
- thường không thể hiện được lý do vì sao (dữ liệu lại như thế này), hay thể hiện được dữ liệu sẽ thay đổi như thế nào trong tương lai,
- phụ thuộc nhiều và ngữ cảnh để có thể suy ra ý nghĩa và ít khả năng dự đoán cho tương lai.

Pattern: mẫu

Pattern is more than simply a relation of relations. It embodies consistency and completeness of relations which, to an extent, creates its own context. Pattern has both an implied repeatability and predictability. ¹

Knowledge: tri thức

"When a pattern relation exists amidst the data and information, the pattern has the *potential* to represent knowledge. It only becomes knowledge, however, when one is able to realize and understand the patterns and their implications.

...

The patterns representing knowledge have a tendency to be more self-contextualizing (i.e. to a great extent, creates its own context).

...

When understood, provides a high level of reliability or predictability as to how the pattern will evolve over time." ¹

"Given these notions [quantitative measures for evaluating patterns], we can consider a *pattern* to be knowledge if it exceeds some interestingness threshold, which is by no means an attempt to define knowledge in the philosophical or even the popular view. As a matter of fact, knowledge in this definition is purely user oriented and domain specific and is determined by whatever functions and thresholds the user chooses." ²

"Một mẫu $E \in L$ được gọi là tri thức nếu như đối với một lớp người sử dụng nào đó, chỉ ra được một ngưỡng $i \in M_i$ mà độ hấp dẫn $I(E, F, C, N, U, S) > i$.

Chú ý rằng định nghĩa trên đây về khái niệm "tri thức" không mang một nghĩa tuyệt đối mà phụ thuộc vào quan điểm của người sử dụng hệ thống KDD [...]

Theo cách hình thức hóa, thuyết minh chính xác cho định nghĩa trên đây về "tri thức" là chọn ngưỡng nào đó $c \in M_c$ (về tính "có giá trị"), $s \in M_s$ (về tính "có thể hiểu được") và $u \in M_u$ (về tính "hữu ích") và khi đó gọi mẫu E là tri thức nếu và chỉ nếu:

$$C(E, F) > c \text{ và } S(E, F) > s \text{ và } U(E, F) > u$$

Wisdom: Trí tuệ

Wisdom arises when one understands the foundational principles responsible for the patterns representing knowledge being what they are. And wisdom, even more so than knowledge, tends to create its own context. ¹

Knowledge Discovery from Database (KDD): Phát hiện tri thức trong cơ sở dữ liệu

"KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. ²

Here, data are a set of facts, and pattern is an expression in some language describing a subset of the data or a model applicable to the subset.

...

Hence, in our usage here, extracting a pattern also designates fitting a model to data; finding structure from data; or, in general, making any high-level description of a set of data.

...

By *nontrivial*, we mean that some search or inference is involved; that is, it is not a straightforward computation of predefined quantities like computing the average value of a set of numbers." ³

"KDD has evolved, and continues to evolve, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in the context of large data sets." ²

"The KDD process involves using the database along with any required selection, preprocessing, subsampling, and transformations of it; applying data-mining methods (algorithms) to enumerate patterns from it; and evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge" ³

"Quá trình KDD thường bao gồm nhiều bước là chuẩn bị dữ liệu, tìm kiếm mẫu, ước lượng tri thức, tinh chế sự tương tác nội tại sau khi chuyển dạng dữ liệu. Quá trình được thừa nhận là không tầm thường theo nghĩa là quá trình đó không chỉ nhiều bước mà còn được thực hiện lặp đi lặp lại, và quan trọng hơn, quá trình đó bao hàm một mức độ tìm kiếm tự động"

4

Quantitative measures for evaluating extracted patterns ² :

- Certainty (e.g., estimated prediction accuracy on new data)
- Utility (e.g., gain in dollars saved because of better predictions or speedup in response time of a system)
- Novelty
- Understandability (e.g., simplicity, the number of bits to describe a pattern)
- Interestingness

Interestingness: độ hấp dẫn

"Độ hấp dẫn: Một tiêu chí quan trọng, được gọi là độ hấp dẫn (interestingness), thường được coi như một độ đo tổng thể về mẫu là sự kết hợp của các tiêu chí giá trị, mới, hữu ích và có thể hiểu được. Một số hệ thống KDD thường sử dụng một hàm hấp dẫn dưới dạng hiển $i = I(E, F, C, N, U, S)$ thực hiện ánh xạ một biểu thức trong L vào một không gian đo được M_i . Một số hệ thống KDD khác lại có thể xác định giá trị hấp dẫn của mẫu một cách trực tiếp thông qua thứ tự của các mẫu được phát hiện.

Trong thực tiễn giải quyết các bài toán khai phá dữ liệu, người ta thường chỉ quan tâm đến độ hấp dẫn, còn các độ đo khác được mặc định coi là thành phần của độ hấp dẫn. Cụ thể là, khi thi hành một loại bài toán phát hiện tri thức cụ thể, một số độ đo tương ứng được tính toán nhằm xác định độ hấp dẫn của tri thức ("mẫu", "luật") đang được xem xét. Chẳng hạn, trong bài toán khai phá luật kết hợp, hai độ đo được xem xét, đó là độ hỗ trợ (xác định phạm vi ảnh hưởng của luật) và độ tin cậy (xác định tính tin cậy của luật) hợp thành độ hấp dẫn của luật kết hợp đã được khai phá. Tương tự, trong bài toán phân lớp, người ta sử dụng hai độ đo cơ bản là độ hồi tưởng (recall - khả năng bao gói ví dụ đúng) và độ chính xác (precision - khả năng chính xác khi xác định ví dụ đúng); đồng thời, một số độ đo mang ý nghĩa kết hợp từ hai độ đo này cũng được sử dụng" ⁴

"An important notion, called *interestingness* (for example, see Silberschatz and Tuzhilin [1995] and Piatetsky-Shapiro and Matheus [1994]), is usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity. Interestingness functions can be defined explicitly or can be manifested implicitly through an ordering placed by the KDD system on the discovered patterns or models." ²

Data Mining: Khai phá dữ liệu

"Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data." ²

"The data-mining component of the KDD process often involves repeated iterative application of particular data-mining methods." ²

"Data mining involves fitting models to, or determining patterns from, observed data. The fitted models play the role of inferred knowledge: Whether the models reflect useful or interesting knowledge is part of the overall, interactive KDD process where subjective human judgment is typically required." ²

Data Warehousing: Kho dữ liệu

"Data warehouses generalize and consolidate data in multidimensional space. The construction of data warehouses involves data cleaning, data integration, and data transformation, and can be viewed as an important preprocessing step for data mining. Moreover, data warehouses provide online analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data generalization and data mining." ⁵







"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process" [^Inm96]

Association Pattern Mining


One of four problems in data mining considered fundamental to the mining process (clustering, classification, association pattern mining, and outlier detection) ⁶

"Phát hiện mối quan hệ kết hợp (associative relation) trong tập dữ liệu là một bài toán quan trọng trong khai phá dữ liệu. Một trong những mối quan hệ kết hợp điển hình là quan hệ kết hợp giữa các biến dữ liệu, trong đó bài toán khai phá luật kết hợp (associative rule) là một bài toán điển hình. Bài toán khai phá luật kết hợp (thuộc lớp phát hiện quan hệ kết hợp), thực hiện việc phát hiện ra mối quan hệ giữa các tập thuộc tính (các tập biến) có dạng $X \rightarrow Y$, trong đó X, Y là hai tập thuộc tính. Về Hình thức, luật kết hợp có dạng giống như phụ thuộc hàm trong CSDL quan hệ, tuy nhiên, nó không được định sẵn từ tri thức miền." ⁴

1. systems-thinking.org, Knowledge Management: <http://www.systems-thinking.org/kmgmt/kmgmt.htm>      

2. Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1-30. Menlo Park, Calif.: AAAI Press.      

3. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. 17, 3 (Mar. 1996), 37. DOI:<https://doi.org/10.1609/aimag.v17i3.1230>  

4. Nguyễn, T. T., Nguyễn, H. N. and Hà Q. T. 2013. Giáo trình Khai phá dữ liệu, Hà Nội: NXB ĐHQGHN   

5. Han, J., Kamber, M. and Pei, J., 2012. *Data Mining: Concepts and Techniques*. 3rd ed. Waltham: Morgan Kaufmann. 

6. Aggarwal, C., 2015. *Data Mining: The Textbook*. Switzerland: Springer International Publishing. 