

Визуализация исследовательской активности организаций с использованием таксономии предметной области

Миркин Б. Г., Насименто С., Мониш-Перейра Л.

mirkin@dcs.bbk.ac.uk

Лондон, Биркбек Колледж и Москва, Высшая школа экономики

В ситуации, когда существует многоуровневая таксономия предметной области, такая как Классификация тематических единиц информатики международной Ассоциации вычислительных машин (АВМ-классификация), деятельность исследовательской организации может быть отображена на эту таксономию для целей анализа и планирования. Эффективность такого отображения зависит от уровня обобщения. Мы предлагаем метод, включающий два этапа обобщения: один через выявление кластеров тематических единиц, второй – через оптимальную «постановку» кластеров в структуре таксономии. Исходные данные формируются в виде матриц связи между предметными единицами на основе информации о том, какие темы развиваются отдельными исследователями. Эта работа поддержана Португальским Фондом науки и техники и опирается на собранные нами данные о некоторых департаментах информатики в Португалии и Великобритании.

Введение: АВМ-классификация и ее использование

Классификация тематических единиц информатики международной Ассоциации вычислительных машин (АВМ-классификация) [1] делит информатику на 11 категорий первого уровня таких как «хардвер», «софтвер», «данные», «теория вычислений», «математика вычислений», «информационные системы», «вычислительные методы», «приложения». Эти категории подразделяются на 81 более мелких предметов второго уровня; из них только 59 не сводятся к тривиальным рубрикам типа «разное». Например, категория «вычислительные методы» включает такие темы как «символическое и алгебраическое манипулирование», «искусственный интеллект», «компьютерная графика», «распознавание образов». Распознавание образов, в свою очередь, делится на единицы третьего уровня: «модели», «кластер-анализ» и пр.

Такие таксономии обычно используются для аннотации и поиска документов или публикаций в различных коллекциях, как это делается для АВМ-классификации на веб-портале Ассоциации вычислительных машин [1]. Некоторые другие применения:

1. Стандарт для автоматически выявляемых онтологий [2];
2. Определение семантического сходства при информационном поиске [3] или электронном обучении [4];
3. Средство ассоциации между потребностями пользователей программного обеспечения и исследователей, создающих новое обеспечение [5].

Мы предлагаем еще одно направление использования АВМ-классификации – для анализа направлений исследований, хотя, конечно, наш метод может быть использован и в других предметных областях. Следует отметить, что существующие практические системы анализа и оценки исследовательской деятельности ориентированы, прежде всего, на анализ индивидуальной активности (см., например, систему РАЭ в Великобритании [8]), тогда как здесь акцент делается на интегральном представлении организации как целого, что может быть полезно для таких задач как

- а Обзор научной деятельности организации.
- б Позиционирование организации в АВМ-классификации.
- в Обзор разработок в стране или иной территориальной единице, с возможностью количественной оценки того насколько достаточен или наоборот избыточен уровень усилий в том или ином направлении.
- г Анализ направлений, не вписывающихся в таксономию, что потенциально может вести к накоплению качества и новым точкам роста или иным неожиданным продвижениям.
- д Планирование инвестиций и структурных изменений.

Метод кластер–постановка

Данная работа включает следующие элементы:

1. разработка электронного интерфейса для того, чтобы каждый член организации мог самостоятельно отобразить тематические единицы, относящиеся к его научным разработкам, и оценить степень интенсивности работы по каждой из них (в принципе, такого

рода информация может быть получена на основе анализа документов в интернете, однако это может применяться только в ситуациях, когда все работы хорошо представлены такими документами);

2. метод вычисления сходства между тематическими единицами;
3. метод для отыскания, возможно пересекающихся, кластеров тематических единиц (экстенциональное обобщение);
4. метод оптимальной постановки кластера тематических единиц в АВМ-классификации (интенциональное обобщение).

Опишем вкратце последние два из них.

Пересекающиеся кластеры

Исходная информация – матрица сходства $A = (a_{ij})$ между тематическими единицами $i, j \in I$, соответствующими висячим вершинам таксономии. Мы используем подход восстановления данных, представленный для случая четких кластеров в [11], а для нечетких – в [10]. Здесь ограничимся только случаем четких кластеров, которые отыскиваются по одному так, чтобы максимизировать отношения Рэлея

$$g(S) = s^T A s / s^T s = a(S) |S|. \quad (1)$$

where

1. $s = (s_i)$ – бинарный индикатор кластера S , $s_i = 1$ для $i \in S$ и $s_i = 0$, в противном случае;
2. $a(S)$ – среднее сходство a_{ij} внутри S , а
3. $|S|$ – число тематических единиц в S .

Квадрат критерия равен доле квадратичного разброса данных, учитываемой кластером S [11].

Критерий локально максимизируется алгоритмом последовательного отбора объектов в (или из) S , начиная с произвольного $i \in I$, ADDI-S [11]. Критерием присоединения или удаления единицы j к S является результат сравнения среднего сходства j и S с адаптивным порогом $\pi = a(S)/2$, выражающего прирост критерия 1. Начиная с каждого $i \in I$, ADDI-S порождает пересекающиеся или даже совпадающие субоптимальные кластеры, которые фильтруются затем так, чтобы образовать базис дизъюнктивной кластерной модели матрицы A .

Оптимальная постановка

Рассмотрим типичный случай, представленный на Рис. 1: кластер тематических единиц (черные висячие вершины) данной организации не совсем

вписывается в структуру таксономии, так как распределен между тремя ее кустами. Ясно, что ему соответствует более общая категория – но ее нет в таксономии. В варианте (А) кластер представлен двумя более общими категориями – ценой введения нескольких пробелов в них и даже одного выброса – в средний куст, не охватываемый выбранными категориями. В варианте (В) – кластер поднят еще выше, до категории, охватывающей все три куста – но со значительно увеличившимся количеством пробелов. Оба варианта – компромиссные, не точно учитывающие реальный кластер, что напоминает задачу о постановке музыкального голоса.

Для формализации этой ситуации введем понятие категории-направления – вершины таксономического дерева, принимаемой в качестве обобщенной характеристики кластера. С каждой категорией-направлением автоматически связывается число пробелов – не членов кластера среди ее детей. Будем штрафовать каждую категорию-направление величиной p , каждый пробел – величиной q , и каждый выброс (элемент кластера, не покрываемый категориями-направлениями) – величиной r . Задача об оптимальной постановке: выбрать такие категории-направления, которые минимизируют суммарную величину штрафа.

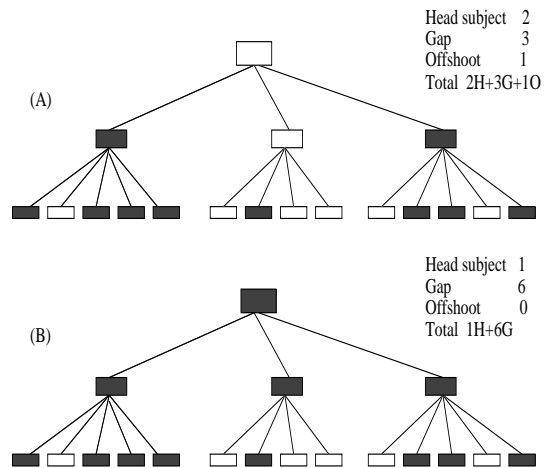


Рис. 1. Варианты постановок кластера в иерархии: постановка (В) более экономна, чем (А), если пробелы значительно дешевле, чем категории-направления.

Решение задачи сожет быть получено с помощью рекурсивного алгоритма вычисления оптимальной постановки в родительской вершине дерева таксономии по оптимальным постановкам

в детях, как это сделано в [9], где для другой прикладной области подобная задача решена в условиях бинарного дерева и при отсутствии выбросов. Особенностью этого алгоритма является необходимость проведения вычислений для каждого из двух различных предположений о природе родительской вершины: (1) категория-направление унаследовано ею от своего родителя; (2) категория-направление не унаследовано, но появилось именно в родительской вершине.

Пример применения

В результате применения изложенного подхода к данным о 49 членах Департамента информатики Нового университета Лиссабона (Португалия). Для простоты используются только показатели сходства между вершинами второго, а не третьего, уровня АВМ-классификации: их оказалось 26 из 59. С помощью алгоритма ADDI-S получено 6 значимых кластеров

1. C11 (вклад 27.08%, интенсивность 2.17), 4 элемента: D3, F1, F3, F4; здесь и далее используются обозначения АВМ-классификации [1].
2. C12 (вклад 17.34%, интенсивность 0.52), 12 элементов: C2, D1, D2, D3, D4, F3, F4, H2, H3, H5, I2, I6;
3. C13 (вклад 5.13%, интенсивность 1.33), 3 элемента: C1, C2, C3;
4. C14 (вклад 4.42%, интенсивность 0.36), 9 элементов: F4, G1, H2, I2, I3, I4, I5, I6, I7;
5. C15 (вклад 4.03%, интенсивность 0.65), 5 элементов: E1, F2, H2, H3, H4;
6. C16 (вклад 4.00%, интенсивность 0.64), 5 элементов: C4, D1, D2, D4, K6.

Оптимальная постановка полученных кластеров, в основном, использует соответствующие категории-направления: F. Теория вычислений (C11), C. Организация вычислительных систем (C13), I. Вычислительные методы (C14), H. Информационные системы (C15), D. Программное обеспечение (C16). Единственное исключение – кластер C12, представляемый двумя категориями-направлениями: D. Программное обеспечение и H. Информационные системы. Это противоречие структуре таксономии, по-видимому, объясняется тем, что в последние годы тема «Инженерия программного обеспечения», охватывающая эти два направления, и имеющая третий ранг в АВМ-классификации, превратилась в дисциплину первого ранга – что следовало

бы учесть в новой структуре. Тот факт, что этот кластер имеет выбросы во все остальные направления, развиваемые в департаменте, может интерпретироваться как определенная его центральность, обеспечивающая его единство.

Заключение

Данный метод может рассматриваться как метод профилирования организаций, в котором обобщение производится для обеих сторон процесса – экстенциональном и интенциональном. Принципиальным является то, что вся работа ведется только в терминах таксономии.

Очевидно, профилирование организации в терминах категорий-направлений может быть сделано более информативным, если специально выделять те из них, которые оказались успешными по тем или иным параметрам (внедрение, цитирование и пр.).

Потенциально данный метод мог бы стать полезным инструментом интеграции и визуализации в анализе деятельности научных и других организаций.

Литература

- [1] The ACM Computing Classification System (1998), url= <http://www.acm.org/class/1998/ccs98.html>.
- [2] C. Thorne, J. Zhu, V. Uren (2005), Extracting domain ontologies with CORDER, Tech. Report kmi-05-14, Open University, 1-15.
- [3] S. Miralaei, A. Ghorbani (2005), Category-based similarity algorithm for semantic similarity in multi-agent information sharing systems, IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology, 242-245.
- [4] L. Yang, M. Ball, V. Bhavsar, H. Boley (2005), Weighted partonomy-taxonomy trees with local similarity measures for semantic buyer-seller match-making, Journal of Business and Technology. Atlantic Academic Press, 1 (1), 42-52.
- [5] M. Feather, T. Menzies, J. Connelly (2003), Matching software practitioner needs to researcher activities, Proc. of the 10th Asia-Pacific Software Engineering Conference (APSEC'03), IEEE, 6.
- [6] S.M. Weiss, N. Indurkha, T. Zhang, F.J. Damerau (2005), Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer Verlag, 237 p.
- [7] S. Middleton, N. Shadbolt, D. Roure (2004), Ontological user representing in recommender systems, ACM Trans. on Inform. Systems, 22(1), 54-88.
- [8] RAE2008: Research Assessment Exercise (2007), url= <http://www.rae.ac.uk/>.

- [9] B. Mirkin, T. Fenner, M. Galperin and E. Koonin (2003), Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes, *BMC Evolutionary Biology*, 3:2
- [10] S. Nascimento, B. Mirkin and F. Moura-Pires (2003), Modeling proportional membership in fuzzy clustering, *IEEE Transactions on Fuzzy Systems*, 11(2), 173-186.
- [11] B. Mirkin (2005), *Clustering for Data Mining: A Data Recovery Approach*, Chapman & Hall /CRC Press, 276 p.