РАЗДЕЛ II. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Нэй Лин

Метод согласования использующие wikipedia как онтология

Курский государственный университет
(Россия, Курск)

doi: 10.18411/sr-10-10-2017-02

idsp: 000001:sr-10-10-2017-02

Аннотация

В статье анализируется один из способов согласования документов и статей использующие Wikipedia как онтология. Определяются подходы к реализации этого способа. Согласование текста традиционными методами осуществляется на основе синтаксической информации, а не семантической информации. Поэтому система кластеризации не понимает значение слов, и при этом в документах имеются синонимы и полисемии. Но здесь существуют и другие проблемы, которые приводят к потере данных и ошибкам в информации. Когда осуществляется замена онтологией одинаковых семантически слов, возникает вероятность потери данных. В этой статье предлагается новый метод согласования, который использует Wikipedia понятия и Wikipedia категории

Ключевые слова: кластеризация, онтология, поиск, семантический вес

Понятие онтологии в информатике используется для представления доменов пространства. Например, в промышленном производстве, научных исследованиях, сельском хозяйстве, военной области и т.д. Посредством онтологии определяются понятия в информатике.

Общая онтология (высший уровень онтологии) представляет собой самую общую онтологию, прямо или косвенно связанную с другими онтологиями. Онтология строится для конкретной предметной области. Например, спорт, медицина, промышленное производство. Задача онтологии представление, обработка и использование знаний предметной областей.

Онтология состоит из трех частей: понятия, свойства и отношения между понятиями. Все они используются для представления и обработки текстов. Использование онтологии для анализа документов заключается в выделении понятий онтологии, которые совпадают с актуальными терминами документа. При этом исходные слова заменяются получающимися понятиями онтологии или добавляется как дополнительные характеристики. После этого отношения между атрибутами и понятиями онтологии используются для анализа документа.

В Wikipedia каждая статья имеет единственный заголовок. Таким образом, имеется сходство Wiki заголовка и понятия в онтологии. Эквивалентность этих понятий позволяет использовать их для перенаправленных ссылок. Это средство структуры Wiki представляет собой иерархическую систему категоризации. Чтобы сравнивать термины текстов и понятия онтологии, надо применять строгие методы согласования. При точном совпадении - метод прямого соответствия актуальных термов и понятий онтологии. Когда актуальные термы не существуют, в понятиях онтологии используют связанные понятия в базе знаний, которые строятся на отношениях между wiki понятиями и документами. После завершения процесса поиска соответствий у каждого документа появляется набор wiki понятий. Поэтому в wiki поисковая система каждый документ связывает с определенной категорией документов.

Wikipedia состоит из множества документов, у которых есть собственные заголовки. Эти заголовки сходны понятиям онтологии. Равнозначные понятия онтологии формируют не прямые ссылки. Поэтому Wikipedia – это иерархическая система

категоризации. Каждая статья связана с одной категорией. Из-за этого структура онтологии становится потенциальной онтологией.

Рисунок(1)- Согласование документов и Wiki понятий

Метод согласования связанных понятий в базе знаний работает за два шагом. На первом шаге строится матрицы Wiki терм - понятия. Каждое слово представляется, как понятия вектор. Связывающие значение TF-IDF[5,4] между Wiki-понятиями и Wiki-термами статьи представляет значение этого вектора. Для того чтобы уменьшить время обработки необходимо убрать незначимые слова с помощью TF-IDF знания.

Wiki Concept \ Wiki Article	Wiki Concept Co				
	CO ₁	CO ₂	CO ₃		CO _N
W ₁	WC ₁₁	WC ₁₂	WC ₁₃		WC _{1N}
W ₂	WC ₂₁	WC ₂₂	WC ₂₃		WC _{2N}
.....					
W _N	WC _{N1}	WC _{N2}	WC _{N3}		WC _{NN}

Рисунок (2)- Матрица Wikistатьи и Wikipонятий

$$r_k^{d_j} = \sum_{w_i \in d_j} tfidf_{d_j}^{w_i} \bullet tfidf_{c_k}^{w_i} \quad (1)$$

Где- d_j – документы в наборе документов

C_k – понятия на Вики понятии

W_i – каждое слово в d_j

На втором шаге надо создать матрицу терм – понятие или документ - понятие, чтобы определить отношения Wiki понятия и документа. При этом используют предыдущую формулу.

1. Hotho, A., Staab, S. and Stumme, G. 2003. Wordnet improves text document clustering. In Proceedings of Semantic Web Workshop, the 26th annual International ACM SIGIR Conference. (Toronto, Canada, Jul. 28-Aug.1, 2003)
2. Hotho, A., Maedche, A. and Staab, S. Text Clustering Based on Good Aggregations, In Proceedings of the 2001 IEEE International Conference on Data Mining. (San Jose, CA, Nov. 29-Dec.02, 2001,). IEEE Computer Society, Washington, DC, 07-608.
3. Yoo, I., Hu, X. and Song, I.-Y. 2006. Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. (Philadelphia, PA, August 20 – 23, 2006). ACM Press, New York, NY, 791 – 796.
4. Zhang, X., Jing, L., Hu, X., et al. A Comparative Study of Ontology Based Term Similarity Measures on Document Clustering. In Proceedings of 12th International conference on Database Systems for Advanced Applications. (Bangkok, Thailand, April 9-12, 2007). 115-126.
5. G. Salton, "The SMART Retrieval System Experiments in Automatic Document Retrieval", New Jersey, Englewood Cliffs: Prentice Hall Inc., 1971.
6. G. Salton and C. Buckley, "Term-Weighting Approach in Automatic Text Retrieval," Information Processing & management, vol. 24, no. 5, 1988, pp. 513-523.