КЛАСТЕРИЗАЦИЯ ДОКУМЕНТОВ НА ОСНОВЕ ОНТОЛОГИИ

CLUSTERING OF DOCUMENTS BASED ON ONTOLOGY

L. Nay

Summary. The article analyzes one of the ways of clustering documents. Approaches to the implementation of this method are determined. Clustering of the text by traditional methods is carried out on the basis of syntactic information, rather than semantic information. Therefore, the clustering system does not understand the meaning of words, and there are synonyms and polysemy in the documents. But there are other problems that lead to data loss and errors in information. When an ontology is replaced by the same semantically word, there is a possibility of data loss. This article proposes a new generalized clustering method that uses Wikipedia concepts and Wikipedia categories.

Keywords: clustering, ontology, search, semantic weight.

Нэй Лин
Аспирант, Курский государственный университет
naylynn16@gmail.com

Аннотация. В статье анализируется один из способов кластеризации документов. Определяются подходы к реализации этого способа. Кластеризация текста традиционными методами осуществляется на основе синтаксический 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

Ключевые слова: кластеризация, онтология, поиск, семантический вес.

Кластеризация текста традиционными методами осуществляется на основе синтаксический 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 обобщенный метод кластеризации, который использует Wikipedia понятия и Wikipedia категории.

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 строиться 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 совпадает с актуальными термами документа. При этом исходные слова заменяются получающимися понятиями онтологии или добавляется как дополнительные характеристики. После этого отношения между атрибутами и понятиями онтологии используется 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 существует, 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.

Xiaohua[7] утверждает, что трудно найти комплексную онтологию, которая представляет всё понятия онтологии. Другой проблемой является то, что замена и добавление свойства происходит достаточно сложно. Когда понятие онтологии заменяет исходное слово, это может привести к потере знаний. Кроме этого, добавле-

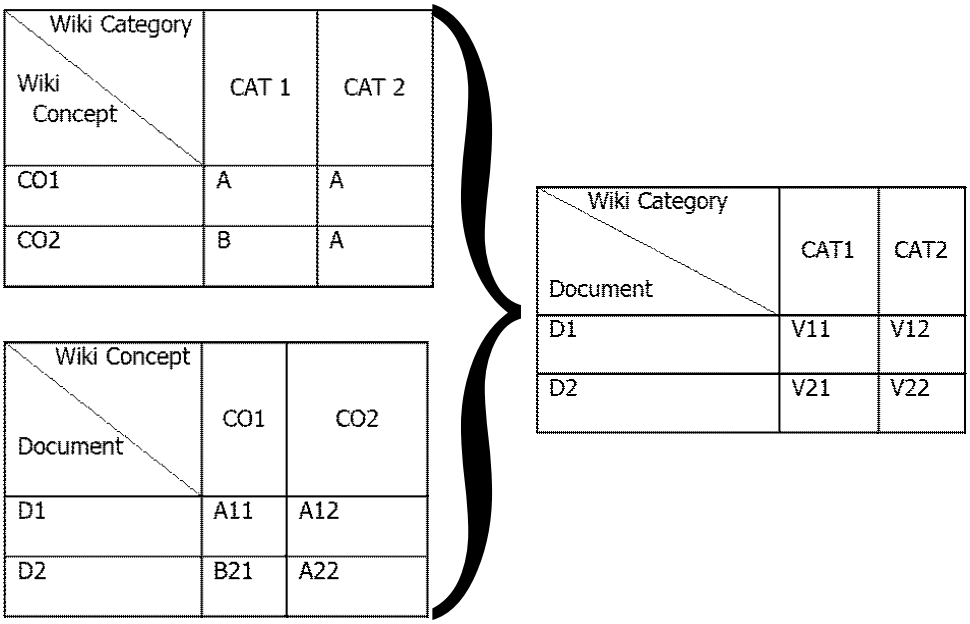

Рис. 1

ние свойства приводит к искажению данных в базе данных.

Wikipedia состоит из множества документов, у которых есть собственные заголовки. Эти заголовки сходны понятиям онтологии. Равнозначные понятия онтологии формируют непрямые ссылки. Поэтому Wikipedia — это иерархическая система категоризации. Каждая статья связанна с одной категорий. Из-за этого структура онтологии становиться потенциальной онтологией.

При группировании документов и представлении документов[7] с использованием понятия онтологии применяются два метода.

Для группирования документов надо создать матрицу сходства между документами и понятиями категории Wikipedia:

- Создать отношения между Wiki понятиями и Wiki категориями,
- Установить вектор соотвествия каждого документа и Wiki понятия,
- Определить совпадающие документы и набор wiki категории.

Для отношения понятия — категории используют отношения между понятиями и категориями, представленными в Wiki, для отношения документ — понятие матрицы используют метод строгого согласования и и метод связанное согласования понятий в базе знаний. Для отношения документы — категории используют отношения понятие — категория и документ — категория.

При использовании метода строгого согласования, вначале анализирует каждый документ и выявляют Wiki понятия. Затем строят вектор из Wiki понятий, комбинируя вики понятия и соответствуюшие документы. Главная проблема точного согласования — синонимы. Как совпадают синонимы фразы и одиночные понятия? Когда Wiki понятия появляются в документах и дают релевантные результаты?

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[5,6] 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.

На втором шаге надо создать матрицу терм — понятие или документ — понятие, чтобы определить отношения Wiki понятия и документа. При этом используют следующуюформулу.

$$r_k^{d_j} = \sum_{w_i \in d_j} tfidf_{d_j}^{w_i} \bullet tfidf_{c_k}^{w_i}$$

Где
d_j — документы в наборе документов;
C_k — понятии на Вики понятии;
W_i — каждое слово в d_j.

Результат вычисления называют связанностью между документами и понятиями. Метод связанности использу-

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}

Рис. 2

ется, когда в Wiki понятии у слов отличные синтаксисы и одинаковые семантики.

После сопоставления документов для каждых наборов документов создает матрица документ — понятия. Применяя эту матрицу и иерархическую концепцию, создается матрица документ — категория. Если использовать точный метод создания отношения документ — понятие, то вместо понятия осуществляем замену соответствующей категорией и затем получаем матрицу категории — документ. Поэтому частота понятия — частота документа. Если бы документ включал более одного понятия, то сумма частот понятий представляет собой частоту категории.

Группирование документов — это автоматическое группирование сходных документов из коллекции документов. Здесь необходимо учитывать объём текстов документов, размер и семантические проблемы. Однозначными в наборе документов считаются слова с одинаковыми характеристиками. Больше всего способов группирования документов объединяют документы по длине и частоте документа. Поэтому характеристика документа зависит от однозначности терма и не представляет дополнительные характеристики.

В хорошем кластере — всё объекты отличаются от других объектов других кластеров. Группированием документов называется обработка данных. Документ кластер можно использовать в базе данных. Подобные документы находятся в подобном кластере. Документы из похожих кластеров извлекаются пользователями. Это является лучшим результатом информационного поиска.

В информационной системе традиционно для пользователя осуществляется сортировка результатов по релевантности запросов. Хотя рейтинг хорошо работает, когда база данных маленькая и запрос корректный, и не будет формировать желаемые результаты, когда база данных крупная и запрос пользователя плохой.

Если пользователи хотят найти подобные документы, то здесь можно выделить интересное свойство, т.к кластеризация способна обнаруживать документы, которые концептуально идентичны, в отличие от поисковых систем, которые могут реализовывать поиск только тогда, когда в документах используются одни и те же слова.

В результате поиска формируется обзор документа. Затем может осуществляться обращение к кластеру релевантности, который выявляет данные интересные пользователям.

Запрос пользователя сравнивается не с отдельными документами, а со с группированными кластерам, поэтому скорость поиска уменьшается. При этом алгоритме используют два метода: иерархический и разделений. Иерархические методы группируют данные объекты, как дерево кластера. Эти методы могут быть далее классифицированы в агломерационную и делительную иерархическую кластеризации в зависимости от того, формируется ли иерархическое разложение восходящим или нисходящим способом.

Чтобы группировать подобные документы надо определить величину различия документов. Здесь считают каждый документ, как кластер и подобные документы объединяются к единому документу. Измерение различия между двумя кластерами может быть реализовано с единой связью, полной связью и средней связью. Когда используется полная связь и средняя связь возникают проблемы. Максимум различия — это расстояние между документами одного кластера и другого кластера. Подобие между документами можно определить с помощью следующей формулы.

$$Sim(d_m, d_n) = sim(d_m, d_n)^{word} + \alpha . sim(d_m, d_n)^{concept} + \beta . sim(d_m, d_n)^{category} ,$$

которая вычисляет подобие между документами α, β, представленными векторами понятий и категорий.

AndresHotho утверждает, что онтология это иерархическая система понятий, поэтому ее можно использовать для кластеризации документов. LeiZhang, Zhichao[8]

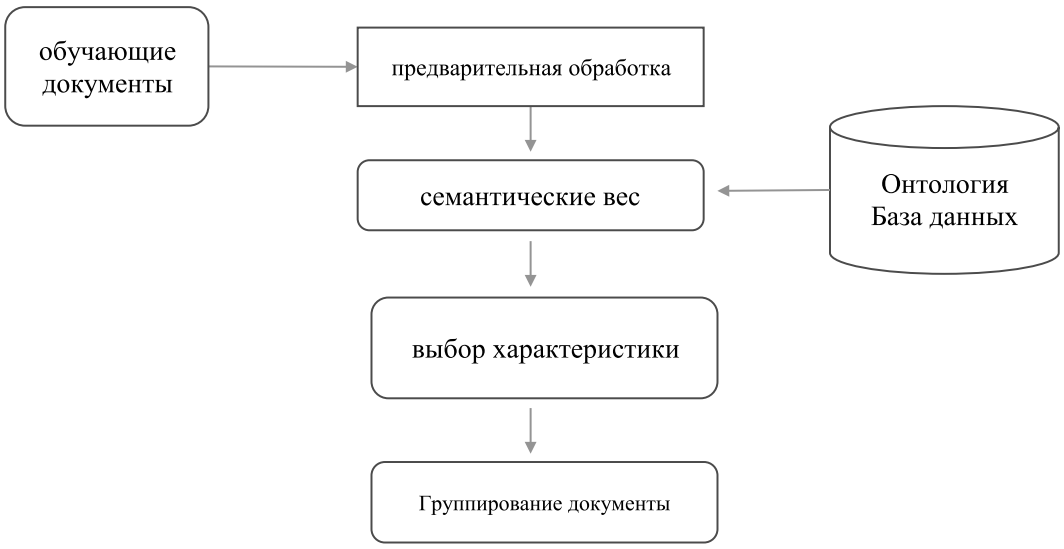

Рис.3. Группирование документов по понятиям

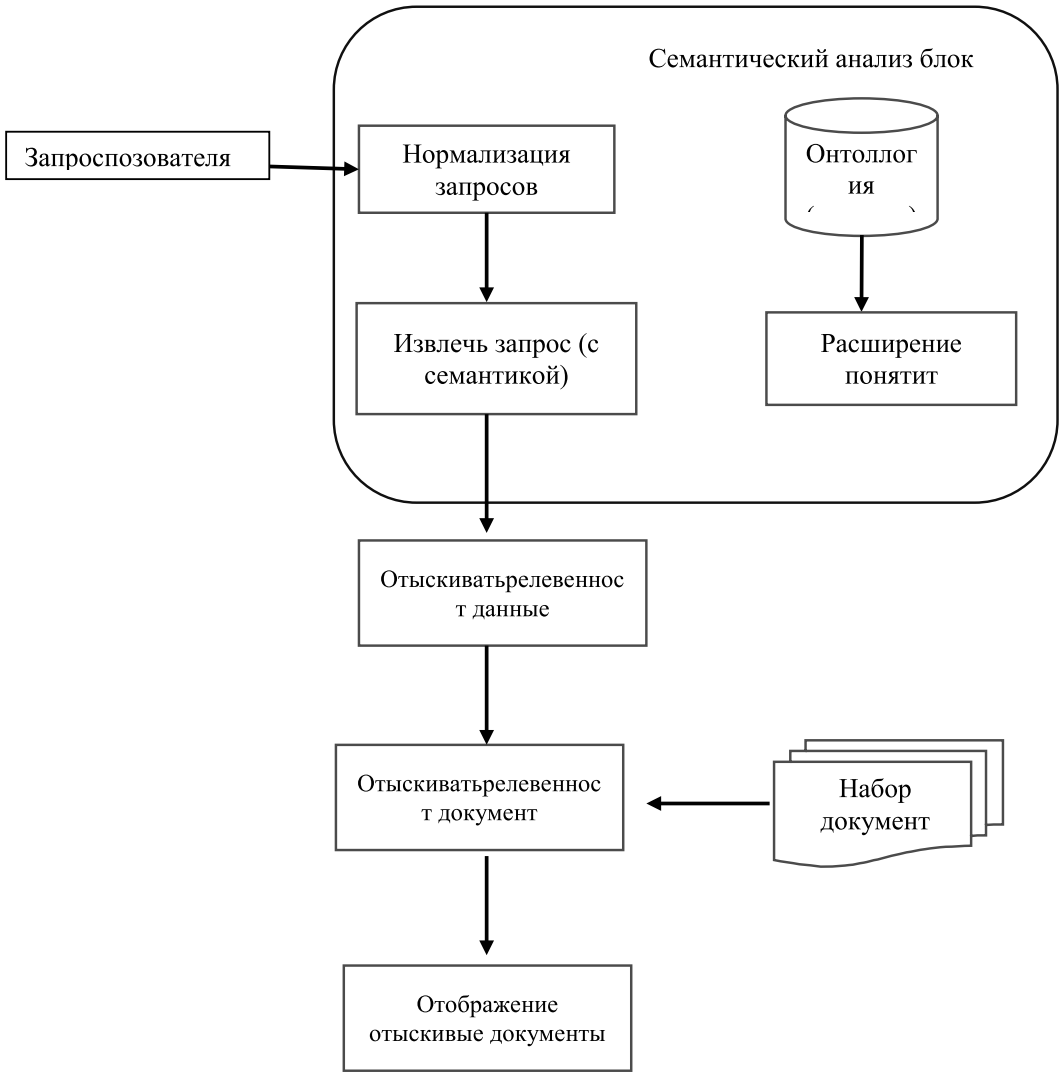

Рис. 4. Семантический информационный поиск

Wang утверждают, что можно использовать алгоритм основанный на кластеризации и характеристику веса. Характеристика веса в дереве онтологии определяет релевантность их характеристики и актуальный терм документа. HmwayHmwayTarandThiThiSoeNyunt утверждает, что определяя вес понятий онтологии, можно группировать документы. Это меняет подход от характеристики к понятию представления. Поэтому кластер документов становиться понятием уровня. Эта система включает 3 части: обработка документы, расчет веса концепции на основе онтологии, определение веса документов кластера. При этом вес — это семантический вес.

Stanislaw Osinski и David Weiss сформулировали алгоритм, который использует частоту фраз для определения метки группирования. Эта метка присваивается результату поиска и фрагменту с этими метками. Ahmed Samehand Amar Kadray [10] добавил к этому алгоритму метод определения частоты фразы. Этой алгоритм может искать частоты синонимов фразы, которые находиться в Wordnet базе данных и с которыми совпадает метка группирования. Использование этих синонимов может объединять документы из разных кластеров. Этой функции не было в первоначальном алгоритме. Её использование увеличивает эффективность кластера. В Wordnet каждому слову поставлены в соответствие части разных синонимных групп и связей между нами. Например — мышь это возможно не только животное в группе синонимов, но и устройство ввода компьютера в электронной группе синонимов.

В английском языке имеется многочисленные синонимы. Чтобы предствлять одиноковые темы в разных документов не существует общих слов. Сансет — это множество сгруппированных синонимов слов[12]. ЯПОКД предлагает вначале искать понятия в документах, потом искать значения понятий используемых алгоритмом априори[13,14], и в конце группировать начальный кластер и выявленные кластеры, которые представляют собой единую частую концепцию. Выявленные начальные кластеры используемым алгоритмом создают дизъюнктный кластер. В результате создается иерархическая древовидная структура. В алгоритм включаются следующие шаги.

- поиск понятии с используемым алгоритмом,
- создание начального кластер для каждого понятия
- создание дизъюнктный кластер, используя функции
- создание кластерного дерева.

ЛИТЕРАТУРА

1. Hotho, A., Staab, S.andStumme, 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
3. International Conference on Data Mining. (San Jose, CA, Nov. 29-Dec.02, 2001,). IEEE Computer Society, Washington, DC, 07–608.
4. 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.
5. 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.
6. G. Salton, «The SMART Retrieval System Experiments in Automatic Document Retrieval», New Jersey, Englewood Cliffs: Prentice Hall Inc., 1971.
7. G. Salton and C. Buckley, «Term-Weighting Approach in Automatic Text Retrieval,» Information Processing & management, vol. 24, no. 5, 1988, pp. 513–523.
8. Xiaohua.Hu, Xiaodan.Zhang, Caimei.Lu, Xiaohua.Zhou, «Exploiting Wikipedia as External Knowledge for Document Clustering», KDD'09, June 28-July 1,
9. L. Jing, M. K. Ng, J. Xu and Z. Huang, «Subspace clustering of text documents with feature weighting k- means algorithm, Proc.of PAKDD, pp. 802–812,2005.
10. HmwayHmway Tar and ThiThiSoeNyunt, «Ontology-Based Concept Weighting for Text Documents», 2011 International Conference on InformationCommunication and Management IACSIT Press, Singapore.
11. Ahmed Sameh, Amar Kadray, «Semantic Web Search Results Clustering Using Lingo and WordNet», InternationalJournal of Research and Reviews in Computer Science (IJRRCS) Vol. 1, No. 2, June 2010.
12. Zeng, Hua-Jun, Qi-Cai He, Zheng Chen, Wei-Ying Ma, and Jinwen Ma. «Learning To Cluster Web Search Results», SIGIR'04, July 2004, Sheffield, South Yorkshire, UK.
13. RekhaBaghel, RenuDhir, «Text Document Clustering Based on Frequent Concepts», 2010 1st International Conference on Parallel, Distributed and Grid Computing (PDGC — 2010).
14. Rakesh Agrawal and Ramakrishnan Srikant, «Fast algorithms for mining association rules». In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proc 20th Int. Conf. Very Large Data Bases, VLDB, pp.487–499, 1994.
15. Imielinski, and A. N. Swami, «Mining Association rules between sets of items in large databases». In Proceedings of ACMSIGMOD International Conference onManagement of Data (SIGMOD93), pp.207–216, Washington, D.C., May 1993.