НАУЧНОЕ ИЗДАНИЕ МГТУ ИМ. Н.Э. БАУМАНА
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС 77 - 48211. Государственная регистрация №0421200025. ISSN 1994-0408
электронный научно-технический журнал

Некоторые методы огрубления графов при оценке релевантности документов

 # 07, июль 2012
DOI: 10.7463/0712.0432649
Вершинин В. Д., Карпенко А. П., Плякин Д. А.
УДК 519.6

Россия, МГТУ им. Н.Э. Баумана
vlad.vershinin@gmail.com
apkarpenko@mail.ru

Введение

Системы поддержки принятия решений (СППР) представляют собой обширный класс интеллектуальных систем, призванных помочь лицу, принимающему решения (ЛПР), в решении слабоструктурированных проблем [1]. Различные варианты классификации СППР представлены, например, в работе [2]. Целям нашей работы соответствует следующая классификация:

- системы, основанные на использовании типовых решений,
- системы, использующие типовые правила синтеза решений,
- системы на основе поиска прецедентов (*Case-Based Reasoning system, CBR-system*).

Работа посвящена актуальной задаче в области искусственного интеллекта - задаче поиска решений в корпоративных базах знаний на основе прецедентов [3]. Работу можно считать продолжением работы [4].

Известны три способа извлечения прецедента или набора прецедентов из базы знаний [3]:

- поиск по прямым указателям в индексе на требуемые признаки,
- поиск по структуре индекса,
- поиск в модели онтологии предметной области.

В работе рассматриваем последний из указанных способов, т.е. полагаем, что речь идет об онтологической базе знаний. Для формализации знаний используем семантическую сеть.

Обычно в *CBR*-системах, построенных на основе семантической сети, в качестве метода извлечения прецедентов используют метод *k*-средних. Однако этот метод неэффективен в тех случаях, когда прецеденты содержат множество несущественных признаков [3]. В работе [4] для извлечения прецедентов из семантической базы знаний предложен метод, основанный на огрублении графа соответствующей семантической сети. Данная работа посвящена оценка эффективности этого метода.

Отметим, что качество решений, полученных с помощью *CBR*-системы, сильно зависит от качества накопленной информации, объёма накопленных данных, объёма знаний о предметной области, а также от способа извлечения прецедентов [3]. Поэтому с самой общей точки зрения целью данной работы является повышение качества решений *CBR*-систем.

Новизна работы заключается в новизне методики и результатов исследования методов ограбления графа, представляющего семантическую сеть предметной области.

В первом разделе работы приведена постановка задачи. Второй раздел представляет исследуемые методы огрубления графа. В третьем разделе изложена методика исследования эффективности указанных методов. Четвертый раздел содержит результаты вычислительных экспериментов и их обсуждение. В заключении сформулированы основные результаты работы и перспективы ее развития.

1. Постановка задачи

Семантическую сеть $S(O)$ рассматриваемой онтологии O представляем в виде взвешенного связного мультиграфа $G(O)$, узлы которого соответствуют концептам $C(O)=\{c_i, i\in[1:n^O]\}$, а ребра – четким бинарным отношениям, каждое из которых принадлежит одному из типов $U_p, p\in[1:m^O]$. Аналогично, семантическую сеть $S(T)$ документа T

определяем в виде связного взвешенного обыкновенного графа $G(T)$, узлы которого соответствуют $n^T \le n^O$ концептам $C(T) \subset C(O)$ документа T, а ребра – связям между ними [4].

Обозначим w_i^O, $i \in [1:n^O]$ веса узлов графа $G(O)$. Пусть также $\{v_{i,j,p}^O, p \in [1:m^O]\}$ - $(1 \times m^O)$-вектор весов ребра (c_i, c_j), $i, j \in [1:n^O]$, $i \ne j$ этого графа. Здесь $v_{i,j,p}^O = 0$, если концепты (c_i, c_j) не связаны между собой отношением типа U_p; $v_{i,j,p}^O = v_p^O$ - в противном случае; v_p^O - вес отношений типа U_p в онтологии O. Алгоритмы определения весов w_i^O, $v_{i,j,p}^O$ рассмотрены в работе [4].

По схеме, предложенной в той же работе, перейдем от взвешенного мультиграфа $G(O)$ к взвешенному обыкновенному графу, в котором вес ребра (c_i, c_j) равен $v_{i,j}^O$. Сохраним за полученным графом прежнее обозначение.

Вес узла графа $G(T)$, соответствующего концепту $c_i \in C(T)$, обозначаем w_i^T, а атрибуты его ребра (c_i, c_j) задаем парой $(l_{i,j}^T; v_{i,j}^T)$, где $l_{i,j}^T$ имеет смысл «расстояния» между концептами c_i, c_j, а $v_{i,j}^T$ - смысл веса этого ребра. В терминах графа $G(T)$ задача построения семантической сети документа $S(T)$ сводится к решению двух следующих задач.

Задача 1 (задача определения топологии графа $G(T)$) - по каким правилам связывать узлы графа $G(T)$ ребрами?

Задача 2 (задача определения весов узлов и атрибутов ребер графа $G(T)$) - исходя из каких соображений, назначать веса w_i^T узлов графа $G(T)$, а также значения атрибутов $l_{i,j}^T$, $v_{i,j}^T$ его ребер?

Рассматриваем методы решения первой из указанных задач, основанные на огрублении соответствующего графа. Метод решения второй задачи предложен в работе [4]

2. Алгоритмы огрубления графа

Суть рассматриваемого подхода к задаче огрубления графа состоит в итерационном стягивании смежных узлов графа G^{α} в узлы графа $G^{\alpha+1}$, где $\alpha = 0,1,2,...$ - номер итерации, $G^{0} = G(O)$ [5]. Специфика нашего случая состоит в том, что запрещено стягивание в один узел тех узлов графа G^{α}, которые принадлежат графу $G(T)$.

Используем алгоритм паросочетаний, когда граф $G^{\alpha+1}$ строится на основе графа G^{α} путем нахождения в графе G^{α} паросочетания и стягивания в мультиузел узлов, входящих в каждую из пар этого паросочетания. При этом непарные узлы графа G^{α} просто копируются в граф $G^{\alpha+1}$. Говоря более строго, используем насыщенные паросочетания, когда хотя бы один узел любого ребра, не вошедшего в паросочетание, инцидентен ребру, вошедшему в паросочетание [5].

Рассматриваем три алгоритма построения насыщенных паросочетаний [6]:

- алгоритм случайных паросочетаний A_{RM},
- алгоритм паросочетаний из тяжелых ребер A_{HEM},
- алгоритм паросочетаний из тяжелых клик A_{HCM}.

Алгоритм A_{RM}. Схема алгоритма для итерации α имеет следующий вид:

1) все узлы C^{α} текущего графа G^{α} объявляем немаркированными;
2) случайным образом выбираем немаркированный узел, еще не включенный в паросочетание (пусть это будет узел c_i^{α});
3) из числа немаркированных узлов, смежных узлу c_i^{α}, случайным образом выбираем узел (пусть это будет узел c_j^{α}), также еще не включенный в паросочетание;

4) если оба узла или один из узлов пары c_i^α, c_j^α не принадлежат графу $G(T)$, то включаем ребро (c_i^α, c_j^α) в паросочетание и узлы c_i^α, c_j^α маркируем;

5) если ни одного немаркированного узла, смежного узлу c_i^α, не существует, то узел c_i^α маркируем и оставляем свободным (чтобы затем перенести его в граф $G^{\alpha+1}$);

6) если в графе G^α имеются еще немаркированные узлы, то переходим к шагу 2.

Алгоритм $A_{НЕМ}$. Схема алгоритма отличается от рассмотренной выше схемы шагом 3, который в данном случае формулируется следующим образом:

- из числа немаркированных узлов, смежных узлу c_i^α, выбираем такой узел c_j^α, также еще не включенный в паросочетание, что вес ребра (c_i^α, c_j^α) является максимальным среди весов всех возможных ребер, связанных с узлом c_i^α.

Алгоритм $A_{НСМ}$. В данном случае также меняется только шаг 3 рассмотренной схемы формирования случайного паросочетания:

- из числа немаркированных узлов, смежных узлу c_i^α, выбираем такой узел c_j^α, также еще не включенный в паросочетание, что реберная плотность $b_{i,j}^\alpha$ мультиузла, который получается стягиванием узлов c_i^α, c_j^α, является максимально возможной по сравнению со всеми иными вариантами выбора узла c_j^α.

Указанную реберную плотность определяем по формуле

$$b_{i,j}^\alpha = \frac{2\left(v_i + v_j + v_{i,j}\right)}{\left(w_i + w_j\right)(w_i + w_j - 1)},$$

где v_i, v_j - суммарные веса ребер, стянутых на предыдущих итерациях в узлы c_i^α, c_j^α соответственно; $v_{i,j}$ - суммарный вес ребра (c_i^α, c_j^α); w_i, w_j - суммарные веса узлов c_i^α, c_j^α соответственно.

Итерации во всех рассматриваемых алгоритмах формирования паросочетаний заканчиваются, когда в результате данной итерации не удалось выделить ни одной пары узлов. Другими словами, итерации заканчиваются, если в текущем графе G^α содержатся только мультиузлы, включающие в себя узлы графа $G(T)$.

Кроме классических алгоритмов A_{RM}, A_{HEM}, A_{HCM}, рассматриваем предложенную нами модификацию этих алгоритмов. Обозначаем модифицированные алгоритмы A_{RM}^M, A_{HEM}^M, A_{HCM}^M соответственно. Модификация учитывает специфику рассматриваемой задачи и призвана повысить эффективность исходных алгоритмов. Суть модификации заключается в том, что на каждой итерации рассматриваем только те мультиузлы графа G^α, которые содержат узлы графа $G(T)$.

Таким образом, схема алгоритма A_{RM}^M, например, имеет следующий вид:

1) все узлы C^α текущего графа G^α объявляем немаркированными;

2) случайным образом выбираем немаркированный узел, который содержит в себе один из узлов графа $G(T)$ и еще не включен в паросочетание (пусть это будет узел c_i^α);

3) из числа немаркированных узлов, смежных узлу c_i^α и не содержащих в себе узел графа $G(T)$, случайным образом выбираем узел (пусть это будет узел c_j^α), также еще не включенный в паросочетание;

4) включаем ребро (c_i^α, c_j^α) в паросочетание и узлы c_i^α, c_j^α маркируем;

5) если ни одного немаркированного узла, смежного узлу c_i^{α}, не существует, то узел c_i^{α} маркируем и оставляем свободным (чтобы затем перенести его в граф $G^{\alpha+1}$);

6) если в графе G^{α} имеются еще немаркированные узлы, то переходим к шагу 2.

3. Методика исследования

Назовем вершину графа $G(O)$, совпадающую с одной из вершин графа $G(T)$, терминальной вершиной.

Все тестовые графы $G(O)$ являются случайными, содержат по $n^O = 250$ узлов и включают в себя пять кластеров. В каждом из кластеров случайным образом выбрана одна из вершин, которой придан смысл терминальной. Таким образом, результирующий граф $G^{\alpha} = G^*$ должен содержать всего пять узлов.

Рассмотрены две группы тестовых графов $G(O)$, параметры которых представлены в таблице 1. Здесь N - номер графа; b^G, b_i^C - реберные плотности графа и его i-го кластера соответственно; min - минимальная реберная плотность, при которой еще не утрачена связность графа; $i \in [1:5]$.

Таблица 1 – Параметры тестовых графов

Группа	$G(O)$	N	b^G	b_1^C	b_2^C	b_3^C	b_4^C	b_5^C
1	G_1	1	min	0,2	0,3	0,5	0,7	0,8
	G_2	2	0,3					
	G_3	3	0,5					
	G_4	4	0,7					
	G_5	5	0,9					
2	G_6	6	0,3	0,1	0,3	0,5	0,7	0,9
	G_7	7	0,5					
	G_8	8	0,7					

Использованы три критерия, определяющих оптимальность рассматриваемых алгоритмов огрубления:

- время решения задачи t, измеренное в CLOCK_TICK с помощью стандартного инструментария языка С,

- реберная плотность b^* результирующего графа,

- величина n^*/n^o, где n^* - число узлов, стянутых в данный мультиузел, а n° - число узлов, которые стянуты в данный мультиузел и при этом принадлежат тому же кластеру, что и соответствующая терминальная вершина.

Легко видеть, что $n^*/n^o \in [0;1]$, и равенство этого отношения единице означает, что соответствующий мультиузел представляет собой стянутый к терминальной вершине кластер.

Для получения статистически достоверных результатов используем метод мультистарта – каждый из представленных в таблице 1 графов генерируем 44 раза и для каждого из полученных графов решаем задачу огрубления. Поэтому, говоря более строго, в качестве первых двух критериев качества алгоритма использованы оценки математического ожидания $\bar{t}$, $\bar{b}^*$ величин t, b_*. В качестве третьего критерия использована величина $\tilde{n}^*$, представляющая собой оценку математического ожидания отношения n^*/n^o, усредненного по всем пяти терминальным вершинам кластера.

4. Вычислительный эксперимент

Интегральные результаты исследования эффективности алгоритмов A_{RM}, A_{HEM}, A_{HCM} и алгоритмов A^M_{RM}, A^M_{HEM}, A^M_{HCM} по критерию $\bar{t}$ представлены на рисунках 1, 2 соответственно. Как и следовало ожидать, рисунки показывают сильную зависимость времени решения задачи от реберной плотности графа b^G. Для данной реберной плотности время решения задачи различными алгоритмами отличатся не более чем на 10%, что лежит в пределах статистической погрешности эксперимента. Другими словами, рисунки показывают близкую эффективность рассматриваемых алгоритмов по критерию $\bar{t}$.

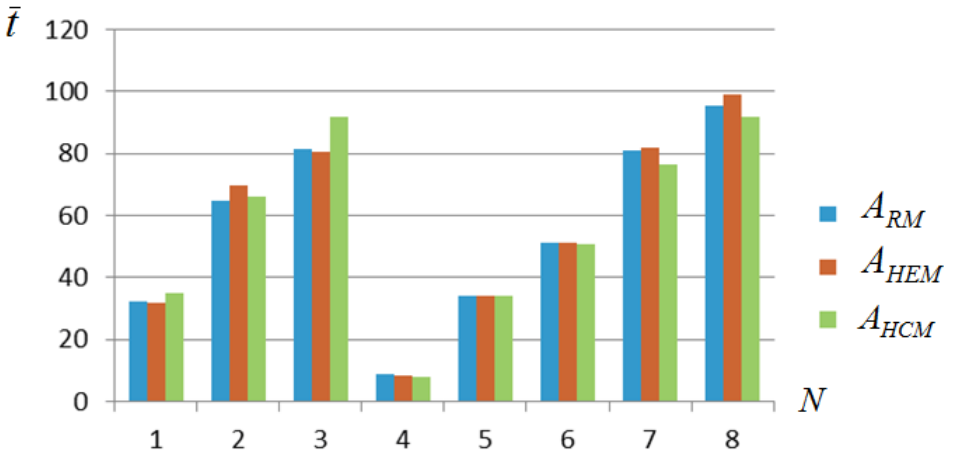


Рисунок 1 – Эффективность алгоритмов A_{RM}, A_{HEM}, A_{HCM} по критерию $\bar{t}$

Значительно более яркими являются результаты сравнительного исследования эффективности алгоритмов A_{RM}, A_{HEM}, A_{HCM} и алгоритмов A_{RM}^{M}, A_{HEM}^{M}, A_{HCM}^{M} (рисунок 3).

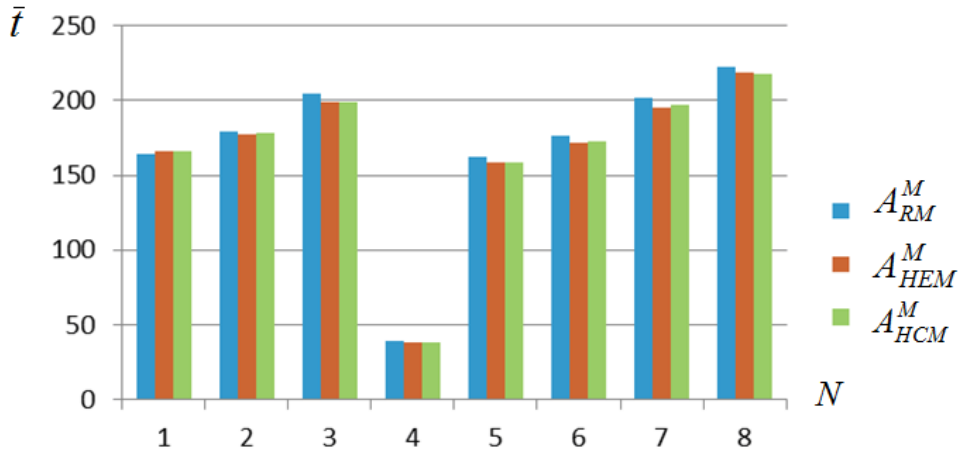
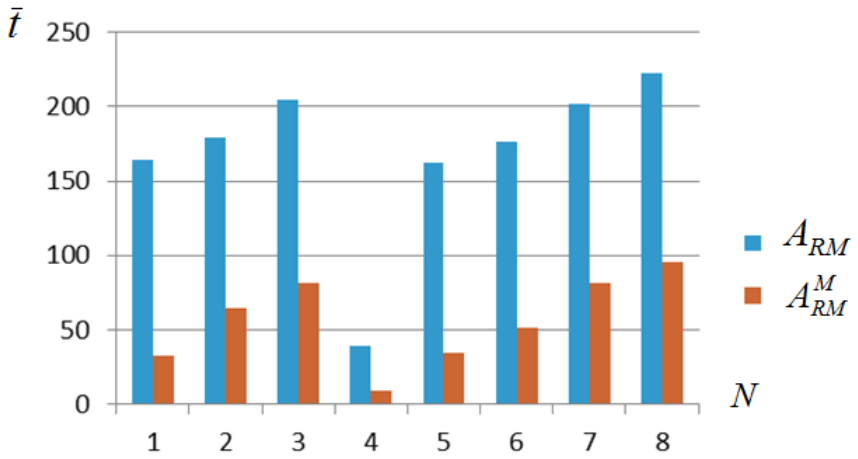
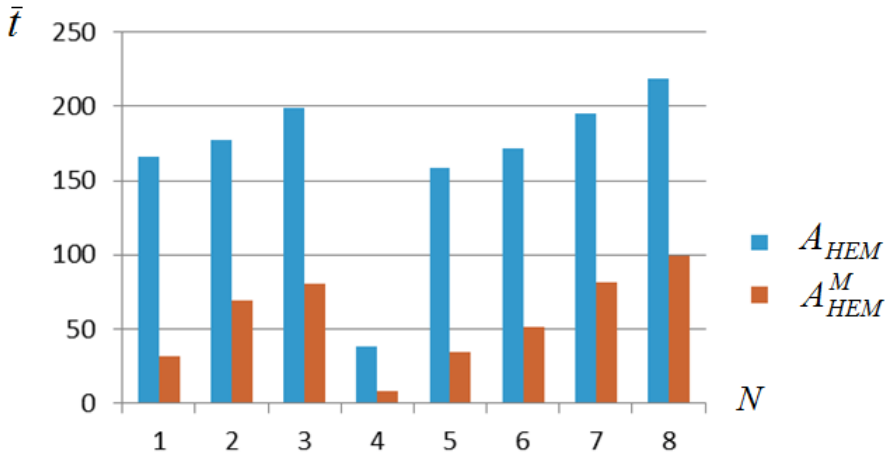


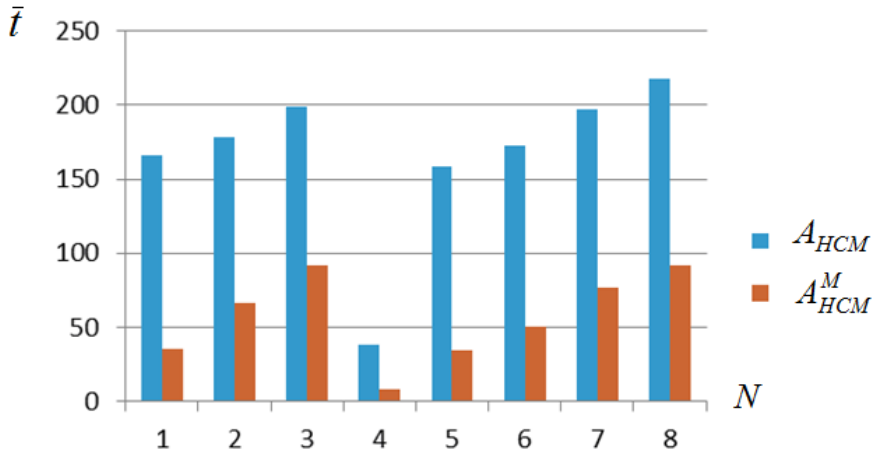
Рисунок 2 – Эффективность алгоритмов A_{RM}^{M}, A_{HEM}^{M}, A_{HCM}^{M} по критерию $\bar{t}$



а) алгоритмы A_{RM}, A_{RM}^{M}



б) алгоритмы A_{HEM}, A^{M}_{HEM}



в) алгоритмы A_{HCM}, A^{M}_{HCM}

Рисунок 3 – Сравнительная эффективность алгоритмов по критерию $\bar{t}$

Рисунки 3а – 3в показывают превосходство модифицированных алгоритмов A^{M}_{RM}, A^{M}_{HEM}, A^{M}_{HCM} над исходными алгоритмами A_{RM}, A_{HEM}, A_{HCM} по критерию $\bar{t}$ от ~100 до ~300%.

Аналогичные результаты исследования по критерию реберной плотности b^{*} результирующего графа представлены на рисунках 4 – 6. Как и в случае критерия $\bar{t}$, рисунки 4, 5 показывают близкую эффективность исходных и модифицированных алгоритмов между собой для всех рассматриваемых классов графов.

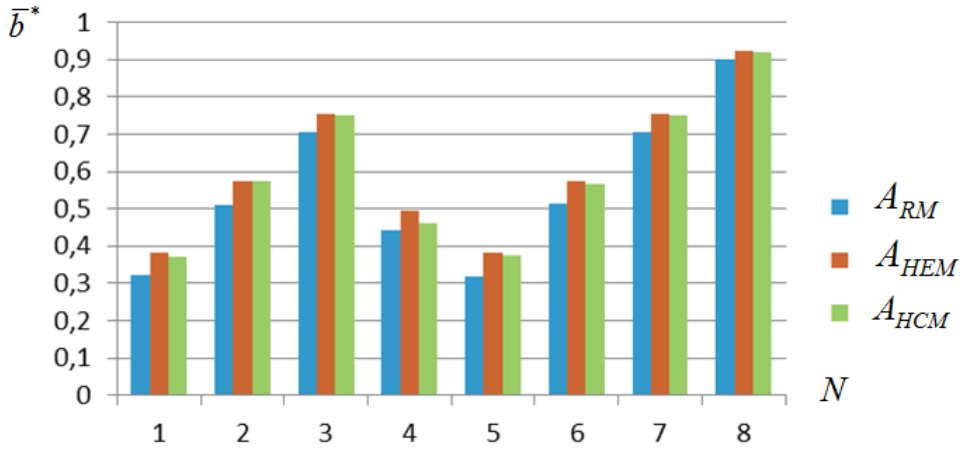


Рисунок 4 – Эффективность алгоритмов A_{RM}, A_{HEM}, A_{HCM} по критерию $\bar{b}^*$

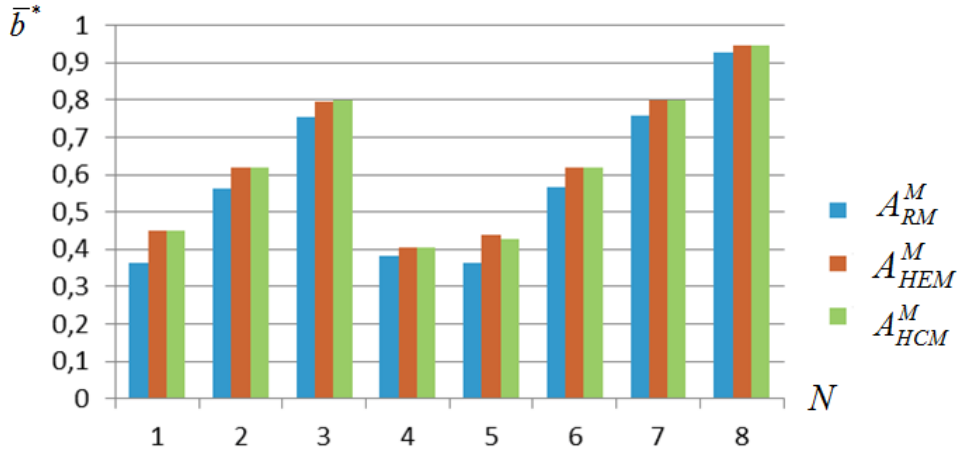
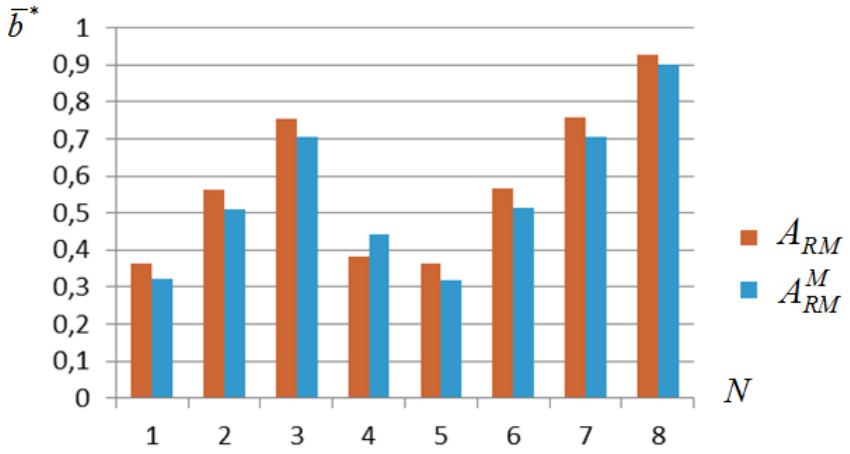
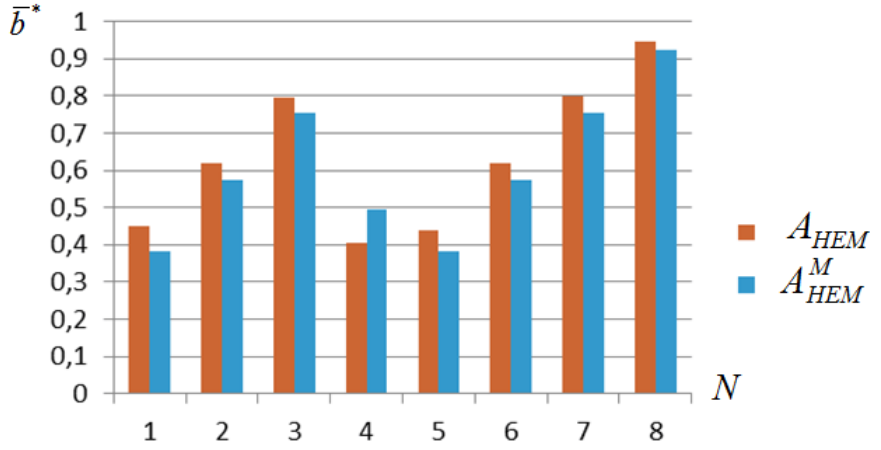


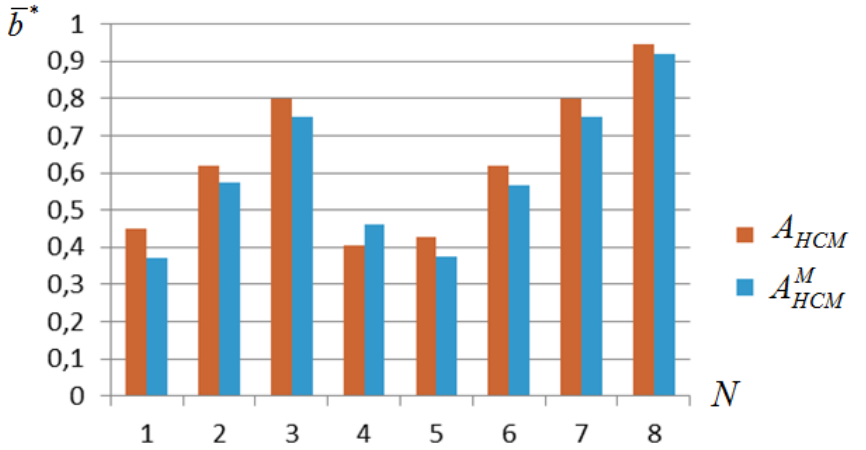
Рисунок 5 – Эффективность алгоритмов A^M_{RM}, A^M_{HEM}, A^M_{HCM} по критерию $\bar{b}^*$



а) алгоритм A_{RM}, A^M_{RM}



б) алгоритмы A_{HEM}, A^M_{HEM}



в) алгоритмы A_{HCM}, A^M_{HCM}

Рисунок 6 – Сравнительная эффективность алгоритмов по критерию $\bar{b}^*$

По критерию реберной плотности результирующего графа рисунки 6а – 6в показывают преимущество модифицированных алгоритмов A^M_{RM}, A^M_{HEM}, A^M_{HCM} над исходными алгоритмами A_{RM}, A_{HEM}, A_{HCM}, достигающее 10%. Заметим, что данный результат относится к случаю малой мощности результирующего графа.

Эффективность исходных алгоритмов A_{RM}, A_{HEM}, A_{HCM} и модифицированных алгоритмов A^M_{RM}, A^M_{HEM}, A^M_{HCM} по критерию $\tilde{n}^*$ иллюстрируют рисунки 7, 8.

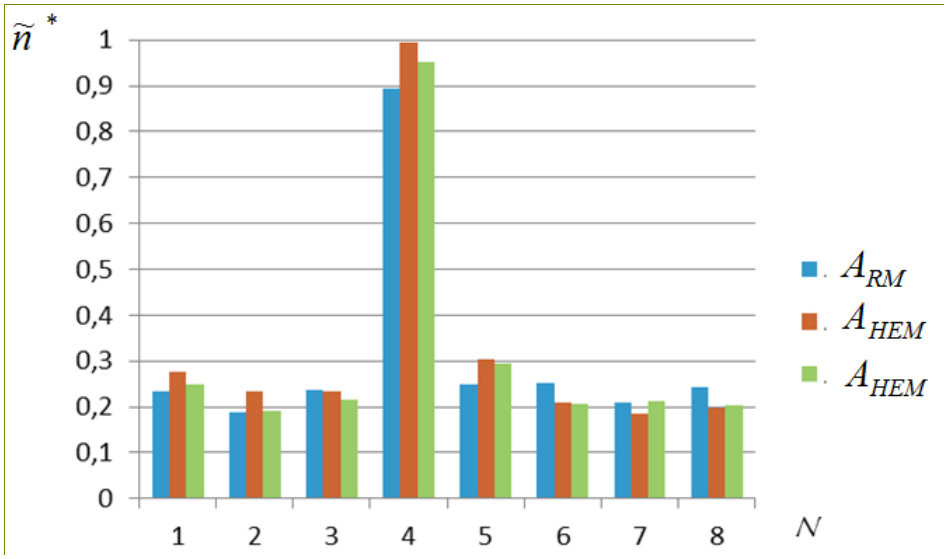


Рисунок 7 – Эффективность алгоритмов A_{RM}, A_{HEM}, A_{HCM} по критерию $\tilde{n}^*$

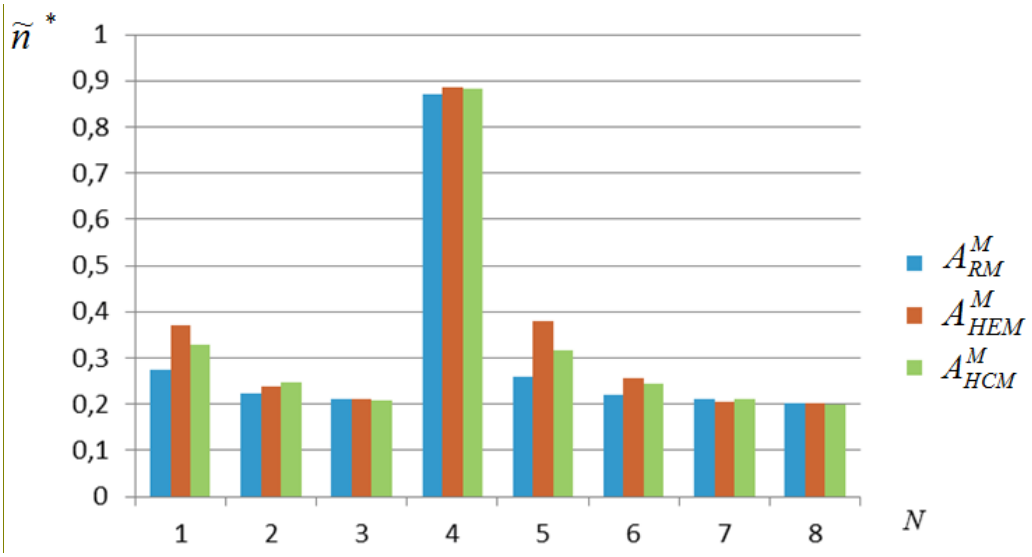
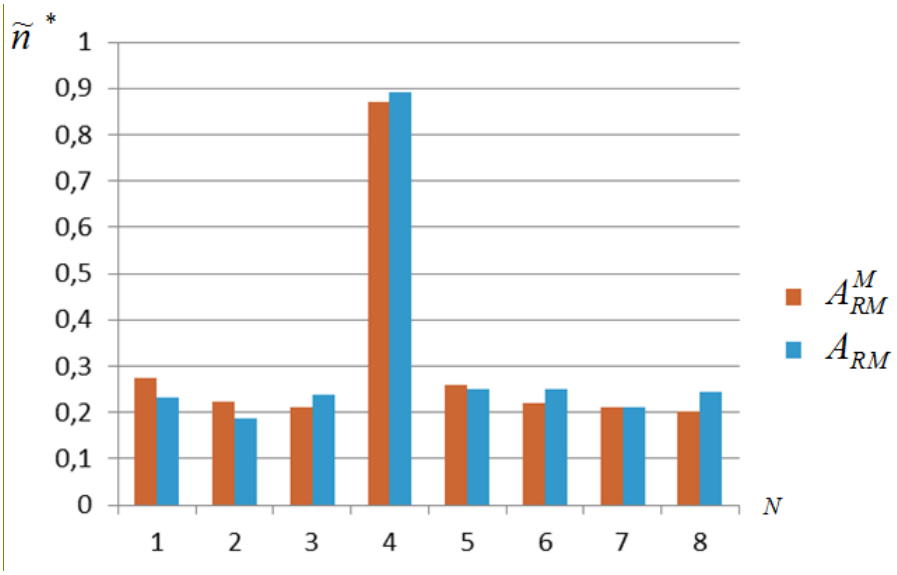


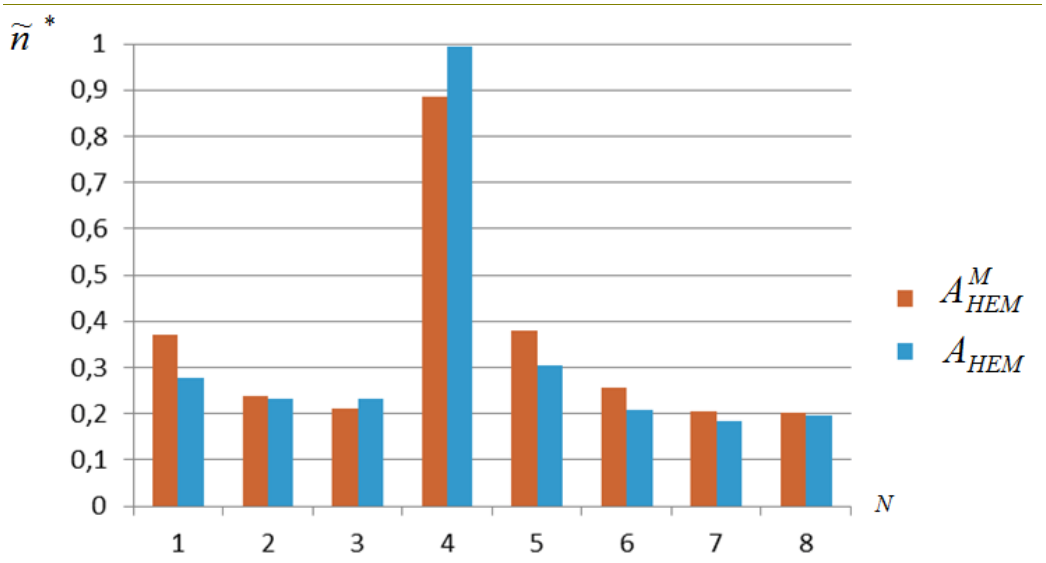
Рисунок 8 – Эффективность алгоритмов A^M_{RM}, A^M_{HEM}, A^M_{HCM} по критерию $\tilde{n}^*$

Рисунки 7, 8 показывают близость соответствующих алгоритмов этих групп между собой по критерию $\tilde{n}^*$ для всех рассматриваемых классов графов и их плотностей.

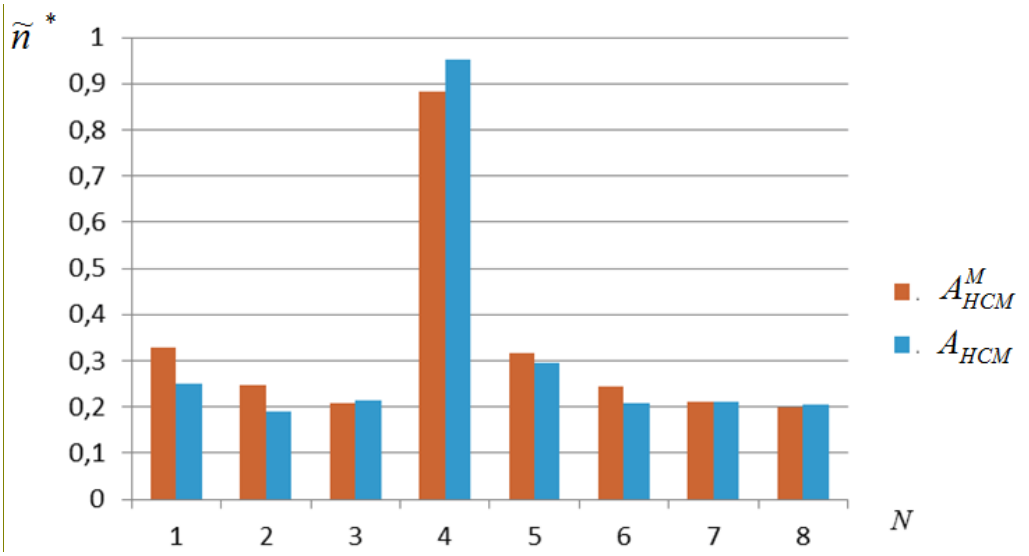
Сравнительную эффективность исходных и модифицированных алгоритмов по критерию $\tilde{n}^*$ иллюстрируют рисунки 9а – 9в.



а) алгоритмы A_{RM}, A_{RM}^M



б) алгоритмы A_{HEM}, A_{HEM}^M



в) алгоритмы A_{HCM}, A^M_{HCM}

Рисунок 9 – Сравнительная эффективность алгоритмов по критерию $\tilde{n}^*$

Рисунки 9а – 9в показывают, что кроме случая слабосвязанных кластеров, модифицированные алгоритмы лучше сохраняют начальную принадлежность вершин кластеру по сравнению с исходными алгоритмами, и это превосходство может достигать 20%.

Заключение

В работе предложена модификация алгоритмов огрубления графа на основе случайных паросочетаний, паросочетаний из тяжелых ребер, также паросочетаний из тяжелых клик. Модификация учитывает специфику рассматриваемой задачи оценки релевантности и ориентирована на повышение эффективности исходных алгоритмов.

Представлена и реализована методика исследования эффективности исходных и модифицированных алгоритмов огрубления графа.

С использованием разработанного программного обеспечения выполнено многокритериальное исследование эффективности алгоритмов. Исследование показало преимущество модифицированных алгоритмов над исходными, достигающее 300%. Результаты исследования предоставляют

лицу, принимающему решения, информацию для многокритериального выбора подходящего алгоритма огрубления графа.

В развитие работы авторы планируют выполнить исследование эффективности указанных алгоритмов для графов $G(O)$, содержащих большее число терминальных вершин.

Литература

1. Ларичев О.И. Теория и методы принятия решений, а также Хроника событий в Волшебных странах. – М.: Университетская книга, Логос, 2006. -292 с.
2. Гайдрик К.В. Системы поддержки принятия решений: эволюция, концепции и некоторые перспективы (http://www.masters.donntu.edu.ua/2010/fknt/sheptulia/library/article05.htm).
3. Норенков И.П. Интеллектуальные технологии на базе онтологий // Информационные технологии, 2010, №1, с. 17-23.
4. Карпенко А.П. Методика оценки релевантности документов онтологической базы знаний // Информационные технологии, 2011, №4, с. 13-23
5. Bui T.N., Chaudhuri S., Leighton F.T., Sipser M. Graph bisection algorithms with good average case behavior // Combinatorica, 1987, N7, pp. 171.191.
6. Бувайло Д.П., Толок В.А. Быстрый высокопроизводительный алгоритм для разделения нерегулярных графов // Вісник Запорізького державного університету, 2002, № 2, с. 1 – 10.

SCIENTIFIC PERIODICAL OF THE BAUMAN MSTU
SCIENCE and EDUCATION
EL № FS 77 - 48211. №0421200025. ISSN 1994-0408
electronic scientific and technical journal

Some methods of graph coarsening in estimating documents relevance

 # 07, July 2012
DOI: 10.7463/0712.0432649

Vershinin V.D., Karpenko A.,P., Plyakin D.A.

Russia, Bauman Moscow State Technical University
vlad.vershinin@gmail.com
apkarpenko@mail.ru

The authors consider the ontology knowledge bases focused on support of decision-making in corporate information systems. The authors believe that search for decision in these knowledge bases is carried out with the use of metadata of the document. Implementation of this approach leads to a graph-coarsening problem. The authors present some methods for solving this problem, and the results of the research of their efficiency.

Publications with keywords: ontology, relevance, decision support, corporate information system, graph-coarsening, semantic network
Publications with words: ontology, relevance, decision support, corporate information system, graph-coarsening, semantic network

References

1. Larichev O.I. *Teoriia i metody priniatiia reshenii, a takzhe Khronika sobytii v Volshebnykh stranakh* [Theory and methods of decision-making, and the Chronicle of events in Magic Lands]. Moscow, Logos Publ., 2006. 29 p.

2. Gaidrik K.V. *Sistemy podderzhki priniatiia reshenii: evoliutsiia, kontseptsii i nekotorye perspektivy* [Decision support systems: Evolution, concepts and some prospects]. Available at: http://www.masters.donntu.edu.ua/2010/fknt/sheptulia/library/article05.htm.

3. Norenkov I.P. Intellektual'nye tekhnologii na baze ontologii [Intellectual technologies on the base of ontologies]. *Informatsionnye tekhnologii*, 2010, no. 1, pp. 17-23.

4. Karpenko A.P. Metodika otsenki relevantnosti dokumentov ontologicheskoi bazy znanii [A method for estimating document relevance in ontology knowledge base]. *Informatsionnye tekhnologii*, 2011, no. 4, pp. 13-23.

5. Bui T.N., Chaudhuri S., Leighton F.T., Sipser M. Graph bisection algorithms with good average case behavior. *Combinatorica*, 1987, no. 7, pp. 171-191.

6. Buvailo D.P., Tolok V.A. Bystryi vysokoproizvoditel'nyi algoritm dlia razdeleniia nereguliarnykh grafov [Rapid high-performance algorithm for the separation of non-regular graphs]. *Visnik Zaporiz'kogo derzhavnogo universitetu* [Herald of the Zaporozhian SU], 2002, no. 2, pp. 1-10.