Модификация модели векторного пространства для ранжирования документов

С.П. Воробьев, М.Б. Хорошко, ЮРГТУ (НПИ), Новочеркасск

В модели векторного пространства документ d и запрос q представляются в виде векторов и релевантность рассчитывается по следующей формуле [1]:

$$score(q,d) = \frac{(\vec{V}(q),\vec{V}(d))}{\|\vec{V}(q)\|\cdot\|\vec{V}(d)\|},$$

Где, $\vec{V}(q)$ – векторное представление запроса, $\vec{V}(d)$ – векторное представление документа. В качестве векторов в эксперименте использовалась оценка веса запроса $w_{t,q}$ и нормированный вес термина в документе - $w_{t,d}$.

$$w_{t,q} = tf * idf,$$

Где tf частота термина в запросе, idf обратная документная частота, вычисляемая по формуле:

$$idf = lg\frac{N}{Df},$$

где N – размер базы документов, Df – количество документов с данным термином.

$$w_{t,d} = \frac{tf}{\sqrt{\sum_{i=1}^{|\text{термин}|} tf^2}},$$

В данном примере вес термина в документе учитывал только частоту термина, но возможны и другие варианты [2] взвешивания документа. Ручной подбор схемы взвешивания для коллекции документов займет большое время, проведем эксперимент для подбора схемы взвешивания используя одну из трех tf, idf, или $tf - idf$ с помощью генетического алгоритма, который получает на вход количество коэффициентов (n) используемых в модели и возвращает подобранные коэффициенты. Общий алгоритм выглядит следующим образом:

1. Создается начальная популяция. Случайным образом из диапазона коэффициентов от C_{min} до C_{max} (диапазон устанавливается для каждого алгоритма), подбираем k_n наборов коэффициентов и переводим их в двоичный вид.
2. Вычисляем приспособленность хромосом. Оцениваем ошибку, для каждого набора коэффициентов.
3. Выбираем двух родителей с наименьшей ошибкой для операции скрещивания.
4. Выбор хромосом для операции мутации.
5. Оценка приспособленности нового набора коэффициентов.
6. Если ошибка n_1 - набора больше заданной ошибки ε_{enter}, то переходим к пункту 3, иначе пункт 7.
7. Полученный набор коэффициентов, который минимизирует ошибку, возвращается в модель поиска.

Рассмотрены более детально основные аспекты:

- Все коэффициенты генерируются изначально случайным образом по равномерному закону при ограничении сверху и снизу. Затем переводятся в двоичный вид, чтобы можно было применять операции скрещивания и мутации.
- Ошибка оценивается по следующей формуле:

$$\varepsilon = \sum_{i=0}^{n} (r(d_i, q_i) - score(d_i, q_i))^2$$

Где, $r(d_i, q_i)$– средняя оценка документа d_i экспертами, по запросу q_i. $score(d_i, q_i)$ – полученная релевантность документа d_i, по запросу q_i.

Эксперимент.Для проверки эффективности применения генетического алгоритма (ГА), сравним полученные метрики оценки для двух систем по 30 запросам.

Полнота ($recall$) вычисляется как отношение найденных релевантных документов к общему количеству релевантных документов:

Полнота характеризует способность системы находить нужные пользователю документы, но не учитывает количество нерелевантных документов, выдаваемых пользователю. Полнота показана на рисунке 1.

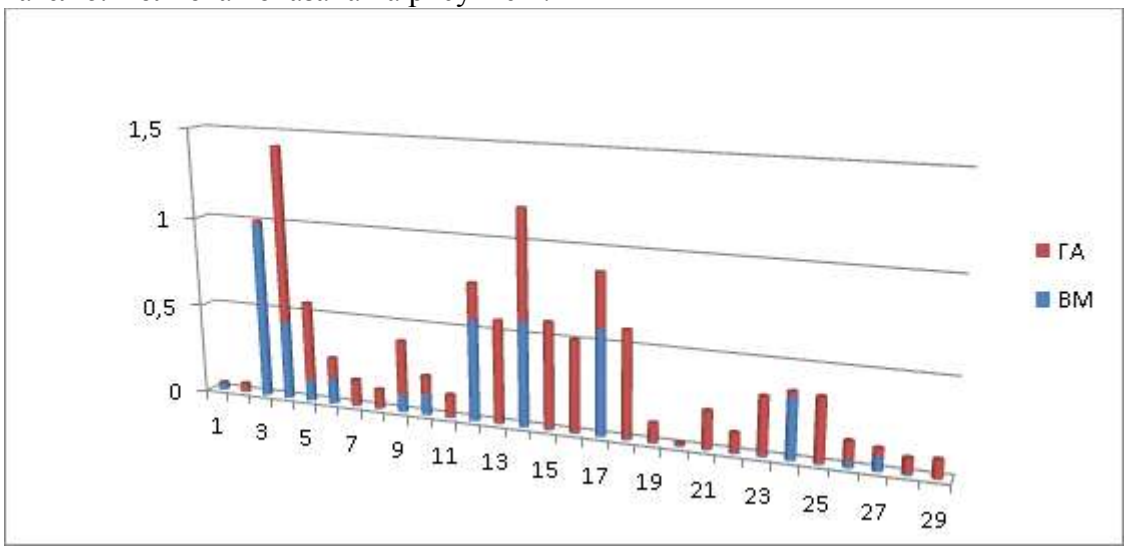

Рис.1. Полнота

В большинстве случаев ГА показывает лучшую полноту. Среднее значение полноты: ГА= 0,245; ВМ=0,153.

Точность ($precision$) вычисляется как отношение найденных релевантных документов к общему количеству найденных документов.

Точность характеризует способность системы выдавать в списке результатов только релевантные документы. Точность алгоритмов показана на рисунке 2.

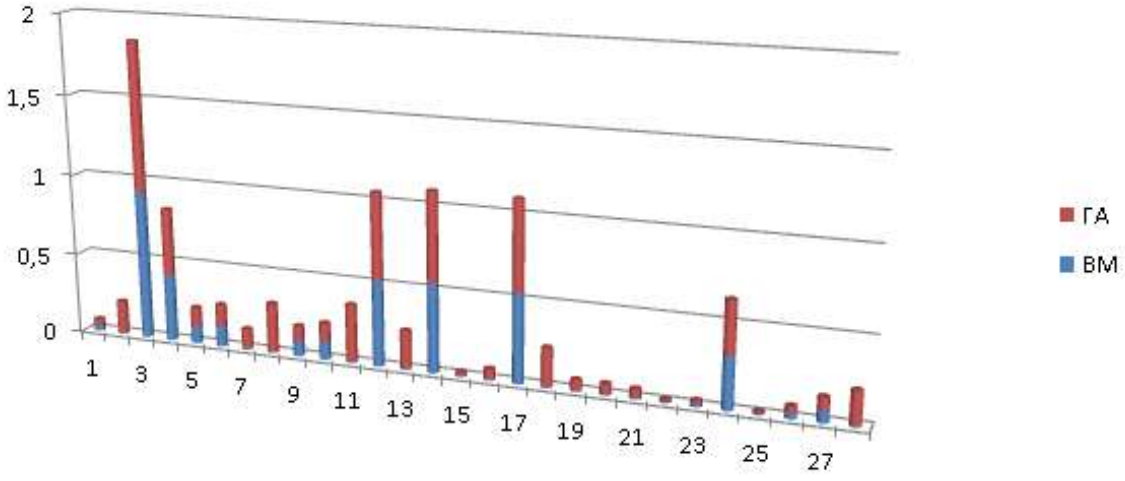

Рис.2. Точность

Среднее значение точности: ГА=0,207; ВМ=0,144.

Аккуратность ($accuracy$) вычисляется как отношение правильно принятых системой решений к общему числу решений. Аккуратность алгоритмов показана на рисунке 3.

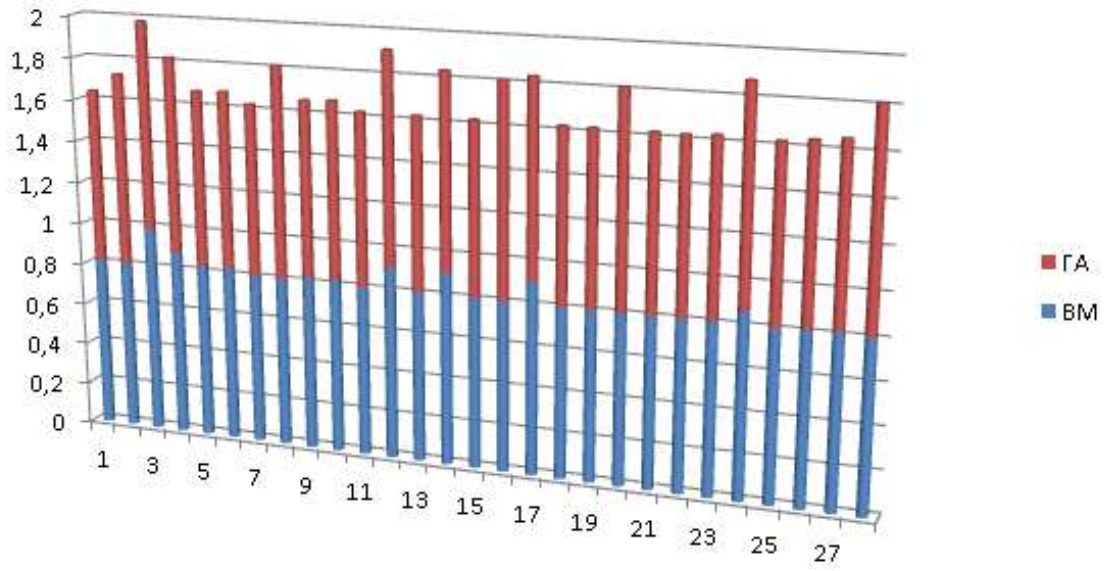

Рис.3. Аккуратность

Среднее значение аккуратности: ГА=0,87; ВМ=0,83.

Ошибка ($error$) вычисляется как отношение неправильно принятых системой решений к общему числу решений. Ошибка алгоритмов полказана на рисунке 4.

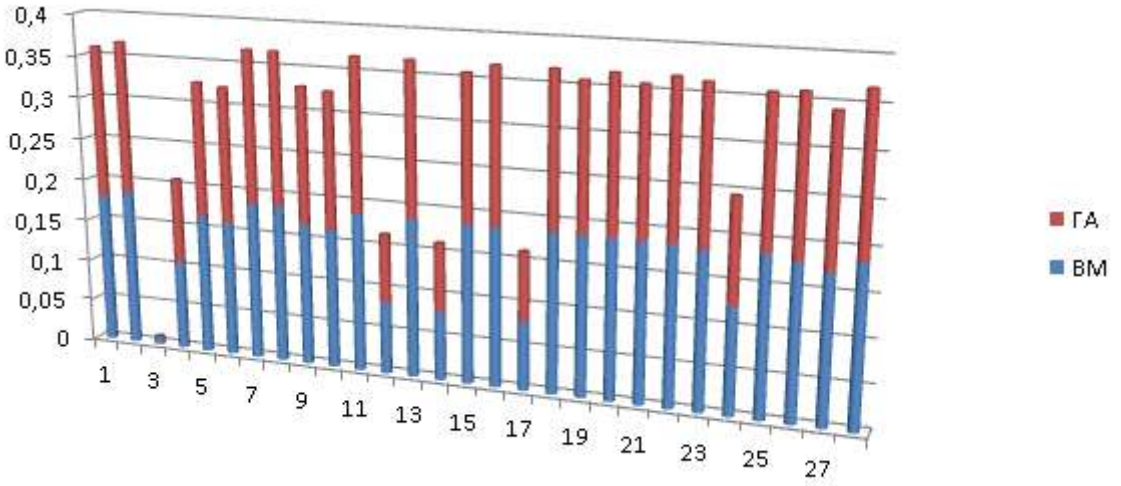

Рис.4. Ошибка

Среднее значение ошибки: ГА=0,153; ВМ=0,16.

F-мера (F) часто используется как единая метрика, объединяющая метрики полноты и точности в одну метрику. F-мера для данного запроса вычисляется по формуле:

$$F = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

Отметим основные свойства:

- $0 \le F \le 1$
- если $\text{recall} = 0$ или $\text{precision} = 0$, то $F = 0$
- если $\text{recall} = \text{precision}$, то $F = \text{recall} = \text{precision}$
- $\min(recall, precision) \le F \le \frac{r+p}{2}$

F-мера алгоритмов полказана на рисунке 5.

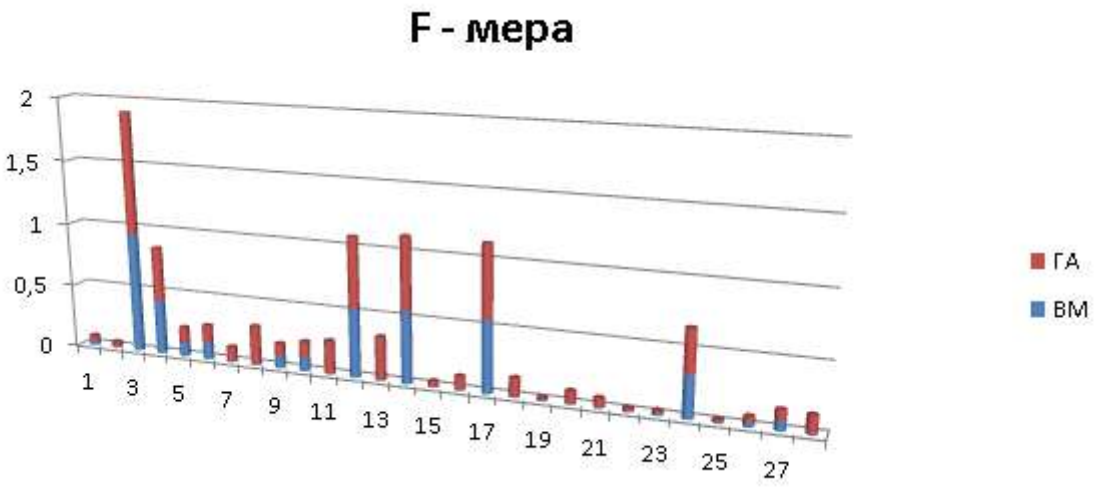

Рис.5. F-мера

Среднее значение f-меры: ГА=0,20; ВМ=0,14.

Таким образом, можно сделать вывод,Модификация с генетическим алгоритмом обладает лучшими значениями метрик, по сравнению с базовым алгоритмом. Но при этом не оправдана сама эффективность использования векторной модели для ранжирования, т.к. вычисление косинусной меры сходства между вектором запроса и каждым вектором документа коллекции, сортировка по релевантности и выбор K лучших документов является довольно затратным процессом и требует выполнения десятков тысяч арифметических операций.

Литература:

1. Маннинг, Кристофер Д. Введение в информационный поиск. М. : Вильямс, 2011.

2. Дубинский А.Г. Некоторые вопросы применения векторной модели представления документов в информационном поиске // Управляющие системы и машины. 2001. № 4.