УДК 004.93'1

АНАЛИЗ КЛАССИФИКАЦИИ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ ВЕСОВЫХ КОЭФФИЦИЕНТОВ

А.М. Луганский[1], С.В. Машталир[2]

[1] ХНУВД, г. Харьков, Украина, alex@adm.univd.kharkov.ua;
[2] ХНУРЭ, г. Харьков, Украина, mashtalir_s@kture.kharkov.ua

Классификация текстов используется для обработки информационных сообщений, для поиска необходимой информации в больших объемах текстов, например, в сети Internet. Рассматривается ряд весовых коэффициентов и проводится анализ классификации множества документов в зависимости от вида запроса и использования различных весовых коэффициентов и их комбинаций.

КЛАССИФИКАЦИЯ, ВЕСОВОЙ КОЭФФИЦИЕНТ, ИНДЕКСАЦИЯ

Введение

Развитие информационных и компьютерных технологий приводит к существенному изменению характера деятельности человека, повышению эффективности и производительности труда в различных областях. Машинная обработка данных облегчает процесс сбора и структурирования информации, позволяет быстрый доступ к ней. Таким образом, все большую роль играют информационно-поисковые системы [1-3].

Основной задачей таких систем является возможность адекватного поиска запрошенных данных за наиболее короткие промежутки времени. В связи с этим возникает целый ряд частных задач, таких как формирование корректного запроса, выделение информативной составляющей, алгоритмы быстрого поиска и т.д. [2, 4, 5]. Вместе с тем важное значение, в частности для библиотечных систем, имеют задачи классификации и систематизации документов [6-11].

Цель классификации текстов — разделение документов на фиксированное число предопределенных категорий, или классов. Каждый документ может попасть во множество категорий, только в одну, либо вообще ни в одну из категорий. Целью использования автоматического обучения является обучение классификаторов на примерах, которые соответствуют данным категориям [7, 10, 11].

Задачу классификации текстов можно интерпретировать по-разному. С математической точки зрения — это задача распознавания образов в алгебраической постановке, а значит, для её решения можно использовать те же принципы, что и для моделирования поведения живых организмов, оценки технологических процессов, распознавания графических объектов и пр. При таком подходе для каждого объекта выделяются векторы признаков. В случае текстов признаками являются слова и взаимосвязанные наборы слов, содержащиеся в тексте [12-14]. В результате обучения информация о соответствии признаков классам текстов сводится в информационную матрицу. Каждый элемент информационной матрицы определяет вес признака при принятии решения о принадлежности признака к данному классу.

Для оценки качества работы поисковой системы применяются два различных типа оценок, основанных на анализе поиска и результатов работы системы. При поиске документов, соответствующих некоторому запросу в поисковой системе, или при классификации документов в рубрицирующей системе (при вынесении решения о соответствии или несоответствии некоторому классу документов) часто используют подход, основанный на весовых коэффициентах [14-16]. Этот этап часто носит название индексации и предназначен для построения классификатора и процесса его обучения. На втором же этапе оценки результата поиска используют метрики [10, 12]. *Цель* работы — анализ возможностей использования весовых коэффициентов для построения систем классификации текстовых данных.

1. Постановка задачи

Пусть имеется некоторое множество документов, для которого необходимо провести классификацию. Существует некоторая система запросов, предоставляемых пользователем для получения информации о правилах классификации. *Задача* заключается в оценке качества полученной в результате проведения запросов информации, а именно — ее соответствия ожиданиям пользователя в плане точности и полноты.

2. Индексация документов с использованием весовых коэффициентов

Использование весовых коэффициентов подразумевает под собой проведение индексации множества документов. При индексации был проведен анализ на множестве документов (100 различных научных статей и книг). Мы использовали метод классификации, называемый ранжированием, то есть множество значений для целевой функции лежит на отрезке [0, 1], и каждый документ с не-

которой долей вероятности относится к тому или иному классу, то есть документ может принадлежать сразу нескольким классам (в отличие от бинарной классификации, в которой есть только два непересекающихся множества). На этапе отнесения документа к классу был использован метод ближайших категорий, то есть документ относится к той категории, к которой он принадлежит с большей степенью вероятности. Это позволяет автоматизировать процесс, в отличие от метода пропорциональности (метод относит документ в разные классы пропорционально степеням вероятности принадлежности документа классу), где может понадобиться частое вмешательство оператора. В методе же ближайших категорий при правильном построении классификатора вмешательство оператора может понадобиться только на этапе обучения.

На первом этапе мы использовали стандартный подход — весовые коэффициенты. В качестве весовых коэффициентов использовались следующие:

α_{12} — коэффициент Луна

$$w_{i,j} = tf_{i,j}$$

(*tf* (term frequency) — частота встречаемости слова в конкретном документе), где $tf_{i,j}$ — количество раз, когда термин i соответствует документу j;

α_{13} — TFIDF

$$w_{i,j} = tf_{i,j} \times \ln \frac{J}{df_i}$$

(TFIDF (IDF = inversed document frequency: насколько редко встречается термин во всех документах) — поиск термина в книге с учетом его встречаемости в книгах коллекции), где J — общее количество документов в коллекции; df_i — количество документов, в которых встречается i термин;

α_{14} — нормализованный TFIDF

$$w_{i,j} = \frac{tf_{i,j}}{ndl_j} \ln \frac{J}{df_i},$$

где $ndl_j = avdl / dl_j$; $avdl$ — средняя длина документа; dl_j — длина j документа, то есть обрабатываемого документа. Нормализация состоит в том, чтобы сумма квадратов всех весов в нем была равна 1;

α_{15} — DTWLM (также носит название LM). Весовой коэффициент, получивший наибольшее применение в статистической модели и выражающийся формулой:

$$w_{i,j} = \frac{tf_{ij}}{dl_j};$$

α_{16} — INQUERY (учет запроса): система INQUERY (предложенная Callan et al. 1992, Allan et al. 1998) базируется на сходстве сетевых документов с терминами, по которым производится индексация, и содержанием запроса (терминами, содержащимися в нем). Таким образом, вероятность соответствия вычисляется по формуле:

$$w_{i,j} = 0{,}4 + 0{,}6 \cdot \frac{tf_{ij}}{tf_{ij} + 0{,}5 + 1{,}5 \cdot \frac{dl_j}{avdl}} \cdot \frac{\ln \frac{J + 0.5}{df_i}}{\ln(J + 1)};$$

α_{17} — метод обратного соответствия (IDF — подсчет количества книг в коллекции, в которых встречается термин) заключается в том, что термины, которые встречаются в очень малом количестве документов, приобретают дополнительный вес (Предложен Спарксом):

$$w_{i,j} = idf_{i,j} = \ln \frac{J}{df_i};$$

α_{18} — BM25, выражающийся в виде:

$$w_{i,j} = \frac{(k_1 + 1) \cdot tf_{i,j}}{K + tf_{i,j}} \cdot \ln \frac{J - df_i + 0{,}5}{df_i + 0{,}5} \cdot (k_2 + 1) \cdot \frac{qtf_i}{k_2 + qtf_i},$$

где qtf_i — количество термина i в запросах; k_1, k_2 — const соответственно равные 2, 5;

$$K = 2 \cdot ((1 - b) + b \cdot \frac{dl_j}{avdl})), \; b = 0{,}75;$$

α_{19} — T-INQUERY, вариация α_{16}, выражающаяся в виде:

$$w_{i,j} = \frac{df_i}{df_i + 200 \cdot ((1 - b) + \frac{dl_j}{avdl})};$$

α_{20} — Okapi коэффициент, выражающийся по формуле

$$w_{i,j} = \frac{(k_1 + 1) \cdot tf_{i,j}}{\frac{dl_j}{avdl} + tf_{i,j}}.$$

Кроме использованных весовых коэффициентов также может использоваться ряд различных коэффициентов: ltn, lnc, dtc, ltc, dtu, Lnu, atn, npn, ntc, BM26 и т.д.

Приведем формулы некоторых из них:

ltn $w_{ij} = (\ln(tf_{ij}) + 1) \cdot idf_j$,

atn $w_{ij} = idf_j \cdot \left[0{,}5 + 0{,}5 \cdot tf_{ij} / \max tf_i\right]$,

lnc $w_{ij} = \frac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik}) + 1)^2}}$,

ntc $w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$,

dtc $w_{ij} = \frac{(\ln(\ln(tf_{ij}) + 1) + 1) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} ((\ln(\ln(tf_{ik}) + 1) + 1) \cdot idf_k)^2}}$,

$$\text{ltc } w_{ij} = \frac{(\ln(tf_{ij})+1)\cdot idf_j}{\sqrt{\sum_{k=1}^{t}((\ln(tf_{ik})+1)\cdot idf_k)^2}},$$

$$\text{dtu } w_{ij} = \frac{(1+\ln(1+\ln(tf_{ij})))\cdot idf_j}{(1-slope)\cdot pivot + slope\cdot nt_i},$$

$$\text{Lnu } w_{ij} = \frac{(\ln(tf_{ij})+1/\ln(1_i/nt_i)+1)}{(1-slope)\cdot pivot + slope\cdot nt_i},$$

$$\text{npn } w_{ij} = tf_{ij}\cdot\ln[\frac{(n-df_j)}{df_j}],$$

$$\text{BM26 } w_{i,j} = \frac{(k_1+1)\cdot tf_{i,j}}{K+tf_{i,j}}\cdot\ln\frac{J-df_i+0{,}5}{df_i+0{,}5}\cdot(k_2+1).$$

Рассмотрим несколько примеров целевого использования коэффициентов. Предположим, необходимо выделить все термины, которые встречаются часто в малом количестве книг, тогда необходимо рассматривать α_{13} и α_{18}. α_{13} должен неограниченно расти $\alpha_{13} \geq 50$, а α_{18} должен быть положительным и иметь значение $20 \leq \alpha_{18} \leq 50$. Если же мы введем еще одно ограничение, например, размер книги должен быть небольшим, то в этом случае дополнительно используются α_{14}, α_{15} и α_{17}. α_{14} должен быть минимальным (в диапазоне от 0 до α_{13}), α_{15} увеличивается и имеет значение $5\cdot10^{-3} \leq \alpha_{15} \leq 5\cdot10^{-2}$, а α_{17} имеет значение больше 1.

Рассмотрим теперь следующий пример. Необходимо выделить термины, которые часто встречаются в большом количестве книг, то есть термины, по которым сложно идентифицировать отдельную книгу, но в то же время это термины, характерные для этого класса книг. В этом случае мы можем использовать α_{18}. Если $\alpha_{18} < 0$, то термин встречается минимум в 50% книг коллекции.

При необходимости использования оценки запроса пользователя нужно брать коэффициенты α_{16}, α_{18}, α_{19}, которые включают встречаемость запрашиваемого термина (терминов) в запросах других пользователей. При этом при одинаковой встречаемости в тексте больший вес α_{16} и меньший α_{19} будет иметь термин, чаще встречающийся в запросах.

Следует отметить, что во многих коэффициентах (α_{14}, α_{16}, α_{18}, α_{20}) используется отношение длины текущего документа и средней длины документов в коллекции. Таким образом, можно учитывать, в какой по объему книге встречается запрашиваемый термин.

3. Оценка точности и полноты по результатам запросов, использующих весовые коэффициенты

Для оценки результатов, представляемых пользователю, существует целый ряд подходов. Одним из них является построение 11-точечного графика полноты/точности, который отражает изменение точности в зависимости от требований к полноте и дает более полную информацию, чем единая метрика в виде одной цифры. При этом под полнотой (recall) понимают отношение найденных релевантных документов к общему количеству релевантных документов:

$$recall = \frac{a}{a+c}.$$

Полнота характеризует способность системы находить нужные пользователю документы, но не учитывает количества нерелевантных документов, выдаваемых пользователю.

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

$$precision = \frac{a}{a+b}.$$

Точность характеризует способность системы выдавать в списке результатов только релевантные документы.

Данный график был построен для коллекции в 50 документов с 6 соответствующими запросам. Система выдает в качестве результатов запроса все эти документы, ранжированные так, что релевантными являются первый, третий, четвертый, пятнадцатый, двадцать первый и тридцать четвертый. Для различных срезов результатов полнота принимает значения 0,17; 0,33; 0,5; 0,66; 0,84 и 1,0. В результате для множества запросов получили результаты, представленные на рис.1.

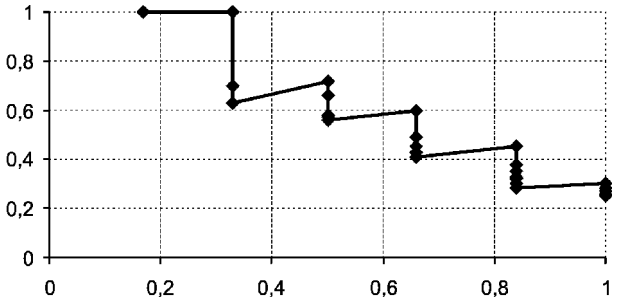

Рис. 1. Кривая полноты/точности для некоторого множества запросов

Как видно из приведенного графика, точность падает с увеличением полноты. Это вполне объяснимо, так как выделить 1-2 релевантных документа из общего числа документов значительно проще, чем 5-6.

Следует отметить, что данный график не в полной мере отражает зависимость точности от полноты, так как не учитывает выбора необходимого весового коэффициента для проведения классификации.

На рис. 2. приведены графики зависимости полноты при использовании в запросах различных коллекций документов.

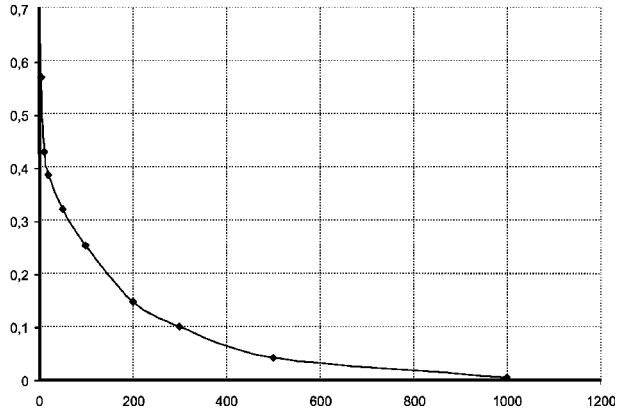

Рис. 2. Зависимость полноты от количества документов

В связи с этим, при поиске большого числа релевантных документов повышается вероятность ошибочного отнесения документа к соответствующим запросу, что неизбежно приводит к снижению точности.

По результатам следует сделать выводы, что полнота падает с увеличением количества документов, используемых для поиска данных по запросу.

На рис. 3 изображена зависимость точности от полноты в зависимости от выбора различных α_i коэффициентов из пункта 3 для проведения классификации. На рис. 3*а* изображены графики для коэффициентов α_{15}, α_{16}, α_{17} соответственно, на рис. 3*б* — α_{19}, α_{18} и α_{20}, на рис. 3*с* — α_{14}, α_{13} и α_{12}.

На основе приведенных графиков мы отмечаем, что на результаты поиска большее влияние оказывает выбор весовых коэффициентов α_{15}, α_{16}, α_{17} и α_{19}, в то время как для остальных точность изменяется более плавно. Однако для коэффициентов α_{20}, α_{13} и α_{12} точность даже при малой полноте оставляет желать лучшего. При этом следует также отметить, что в данном случае мы не рассматривали зависимость наших результатов от количества ключевых терминов или фраз, задаваемых пользователем. Однако данный параметр может существенно повлиять на изменение точности и особенно полноты получаемых в результате данных.

Далее необходимо выяснить, как влияет совместное использование коэффициентов на точность и полноту классификации множества документов. Эти зависимости приведены на рис. 4.

Из приведенных зависимостей можно сделать вывод о целесообразности совместного использования коэффициентов, так как точность при той же полноте повышается примерно на 10%. Однако необходимо учитывать то, что некоторые коэффициенты несовместимы между собой и их использование может привести к снижению точности классификации по сравнению со значением каждого из них, как это представлено на графике ■. В этом случае, несмотря на достаточно высокое значение точности при малой полноте, с ростом полноты точность резко падает, что в целом является нежелательным. Также следует отметить, что увеличение числа коэффициентов приводит к большим вычислительным затратам, а следовательно — к увеличению времени классификации.

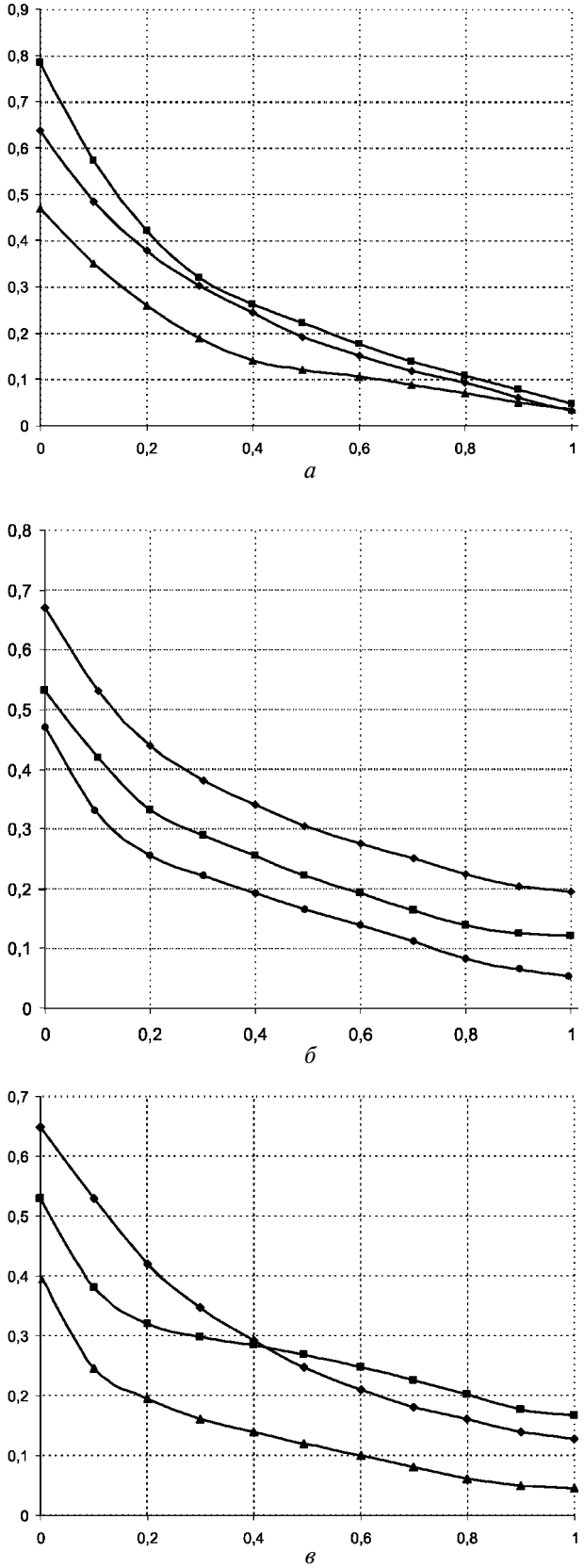

Рис. 3. Зависимость точности от полноты при использовании различных α_i

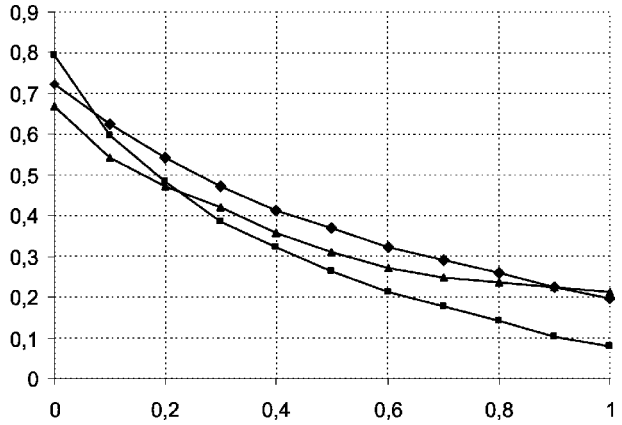

Рис. 4. Отношение точности к полноте при совместном использовании весовых коэффициентов

Результаты и перспективы исследований

Необходимо отметить, что использование весовых коэффициентов на этапе индексации позволяет создавать качественные классификаторы. Из приведенного анализа точности и полноты результатов запросов можно сделать вывод о целесообразности использования комбинированных весовых коэффициентов с целью создания более устойчивых к входным данным запросов классификаторов. Это в первую очередь связано с тем, что разные коэффициенты могут дополнять друг друга, что позволяет увеличить полноту и точность, так как будет находиться большее число документов, релевантных запросу. Однако еще следует более широко исследовать вопрос возможного количества используемых одновременно весовых коэффициентов, то есть как влияет на полноту и точность использование одновременно n коэффициентов. Ведь может получиться, что рост вычислительных затрат не соответствует росту значений точности и полноты, или же вообще мы придем к ситуации, когда увеличение количества коэффициентов не сказывается положительно на значениях точности и полноты.

Следует отметить, что полученные результаты позволяют рассчитывать на эффективное использование комбинированных весовых коэффициентов при создании классификаторов для различных поисковых систем и в электронных библиотеках.

Список литературы: 1. *Chen H., Houston A., Sewell R.., and Schatz R.* Internet browsing and searching: User evaluations of category map and concept space techniques // Journal of the American Society for Information Science. — Vol. 49, № 7. – 1998. — P. 604-618. 2. *Gordon M., Pathak P.* Finding information on the World Wide Web: The retrieval effectiveness of search engines // Information Processing and Management. — Vol. 35, №2. — 1999. — P. 141–180. 3. *Menczer F.* Complementing search engines with online Web mining agents // Decision Support Systems. — Vol. 35, № 2. — 2003. — P. 195–212. 4. *Харламов А.* Автоматический структурный анализ текстов // Открытые системы. — № 10. – 2002. 5. *Aslam J.A., Pavlu V. and Savell R.* A unified model for metasearch, pooling, and system evaluation. In: Proceedings of theTwelfth International Conference on Information and Knowledge Management.NewOrleans, LA, — 2003. — P. 484–491. 6. Advances in automatic text summarization / Mani I. and Maybury M.T. (eds.) Cambridge, Massachusetts: MIT Press. — 1999. — 442 p. 7. *Berry M.W.* Survey of text mining: clustering, classification, and retrieval. New York: Springer-Verlag. — 2003. — 244 p. 8. *Wai L., Ruiz M., and Srinivasan P.* Automatic Text Categorization and Its Application to Text Retrieval // IEEE Transactions on Knowledge and Data Engineering. — Vol.11. — 1999. — P. 865–879. 9. *Willett I.* Query specific automatic document classification / International Forum on Information and Documentation, — Vol. 10. — 1985. — P. 28-32. 10. *Goldstein J., Kantrowitz M., Mittal V. and Carbonell J.* Summarizing text documents: Sentence selection and evaluation metrics / In Proceedings of SIGIR, — 1999. — P.121-128. 11. *Алешин Л.И.* Классификация информационных ресурсов электронных библиотек: По результатам поиска в Интернете // Библиография. — 2003. — №4. — С. 3-7. 12. *Sandor D.* Mathematical Foundations of Information Retrieval. — Kluwer. 2001. — 304 p. 13. *Лоуренс С.* Контекст при поиске в Web // Открытые системы. — №12. — 2000. — С. 62–66. 14. *Jacso P.* Citation searching // Online Information Review. — Vol. 28, No 6. — 2004. — P. 454-460. 15. *Воройский Ф.С.* Индексирование документов в автоматизированных библиотечно-информационных системах // Библиотека. — 1996. — №9. — С.42–44. 16. *Cleveland D.B. and Cleveland A.D.* Introduction to Indexing and Abstracting, 3rd ed. Libraries Unlimited, Englewood, CO. — 2000. — 283 p.

Поступила в редколлегию 18.10.2007