ОӘЖ 004.042

АЛГОРИТМ ҚҰРУ ЖӘНЕ ОНЫҢ ҒЫЛЫМИ МӘТІНДЕРДІ ТАЛДАУДАҒЫ ӘСЕРІ

Тұрарбек Әмина Мұратбекқызы

2 курс магистранты**,** «Ақпараттық технологиялар» кафедрасы, Технологиялық факультет, Қ.Құлажанов атындағы қазақ технология және бизнес университеті

Ғылыми жетекшілер: Муханова А.А., PhD, acc. профессор, Алтынбек С.А. PhD

Бұл мақалада Байес тәсілін қолдана отырып, ғылыми мәтіндерді талдау алгоритмін жасау қарастырылады. Ғылыми жарияланымдар санының қарқынды өсуі жағдайында ақпаратты тиімді алу өте маңызды болады. Жұмыста мәтіндерді өңдеудің қолданыстағы әдістері, соның ішінде лексикалық, Машиналық оқыту және Байес модельдері талданды, олардың артықшылықтары мен кемшіліктері анықталды. Ұсынылған алгоритм талдаудың жоғары дәлдігін қамтамасыз ете отырып, ғылыми деректерді жіктеу және құрылымдау үшін Байес ықтималдығын пайдаланады. Алгоритм – ғылыми мәтіндерді өңдеудің тиімділігін арттыруға, қателерді азайтуға және оларды жіктеу процесін автоматтандыруға қабілетті. Болашақта терең нейрондық желілерді біріктіру және модельдің көп тілді корпустарға мүмкіндіктерін кеңейту жоспарлануда.

Кілт сөздері: Байес тәсілі, ғылыми мәтіндерді талдау, машиналық оқыту, деректерді жіктеу, мәтінді өңдеу, алгоритм, нейрондық желілер, автоматтандыру, лексикалық талдау, ықтималдық модельдері.

Кіріспе

Ақпараттық шамадан тыс жүктеменің қазіргі дәуірінде ғылыми мәтіндердің үлкен көлемін талдау барған сайын күрделі міндетке айналуда. Әртүрлі салалардағы ғылыми жарияланымдар санының қарқынды өсуі жағдайында қажетті ақпаратты тиімді алу ғалымдар мен зерттеушілер үшін өте маңызды. Мәтінді талдаудың дәстүрлі әдістері көбінесе жылдамдық пен дәлдік тұрғысынан жеткіліксіз болып шығады, бұл жетілдірілген құралдар мен әдістерді қолдану қажеттілігін көрсетеді.

[1] Осындай шешімдердің бірі-мәтінді талдау процесін жақсартуға арналған арнайы алгоритмдерді әзірлеу. Алгоритмдер, әсіресе Байес теоремасы сияқты ықтималдық модельдеріне негізделген, ғылыми мәтіндерден алынған

ақпараттың дәлдігі мен өзектілігін едәуір арттыра алады. Талдаудың негізгі аспектілерін автоматтандыру арқылы бұл алгоритмдер қол еңбегін азайтады, қателерді азайтады және зерттеушілерге деректерді өңдеуге емес, тереңірек түсінуге назар аударуға мүмкіндік береді.

Бұл мақалада ғылыми әдебиеттерді талдауға арналған алгоритм құру, оның құрылымы, функционалдығы және мәтіндік талдаудың тиімділігі мен сапасына әсері қарастырылады. Сондай-ақ, тиімді шешімдер қабылдауға және білімді ашуға ықпал ету үшін мұндай алгоритмдерді зерттеудің жұмыс процестеріне қалай біріктіруге болатындығы қарастырылады.

Әдістері

Бұл зерттеудің негізгі мақсаты ғылыми мәтіндерді талдау үшін ықтималдық теориясына

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

атап айтқанда *Байес теоремасына* негізделген алгоритмді әзірлеу және бағалау болып табылады. Мұнда $P(A|B)$ – B сөздерін ескере отырып, мәтіннің A санатына жату ықтималдығы. $P(B|A)$ - A санатындағы B сөздерінің пайда болу ықтималдығы. $P(A)$ – A санатының априорлық ықтималдығы. $P(B)$ - Мәтіндегі B сөздерінің пайда болу ықтималдығы (еленбейді, өйткені барлық санаттар үшін бірдей).

Төменде алгоритмді әзірлеу және тестілеу кезінде қолданылатын әдістеме сипатталған:

Деректерді жинау: [3], [4] Талдау үшін әртүрлі салалардағы ғылыми мақалалар туралы көптеген мәліметтер жиналды. Бұл дерекқорға құрылымдық метадеректер (мысалы, атаулар, кілт сөздер) және құрылымдалмаған мазмұн (мысалы, аннотациялар, толық мәтіндер) кіреді. Бұл мәліметтер жиынтығы алгоритмнің ғылыми мәтіндерді жіктеу, саралау және жалпылау қабілетін тексеруге негіз болды.

Алгоритм дизайны: [2] Алгоритм Байес ықтималдығы теориясына негізделген, бұл берілген ғылыми мәтіндегі белгілі бір белгілердің (мысалы, кілт сөздер, сөз тіркестері) өзектілігін болжауға мүмкіндік береді. Даму барысында сөз жиілігі, терминдердің маңыздылығы және контекстік өзектілік сияқты негізгі параметрлер анықталды және алгоритмге біріктірілді. Бұл алгоритмге мәтіннің әртүрлі бөліктерінің жалпы тақырыпқа немесе зерттеу сұрағына үлес қосу ықтималдығын бағалауға мүмкіндік береді.

Іске асыру

Алгоритм Python және әртүрлі табиғи тілдерді өңдеу кітапханалары (natural language processing - NLP) арқылы жасалған. Негізгі функцияларға мәтіннен мағыналы үлгілерді алуға мүмкіндік беретін токенизация, стемминг және жиілікті талдау кіреді. Байес теоремасы сәйкес және маңызды емес мәтіндердегі

терминдер мен сөз тіркестерінің пайда болуының шартты ықтималдықтарын есептеу үшін жүзеге асырылды.

Тестілеу және валидация

Алгоритм сәйкес мәтіндерді анықтаудағы дәлдігін бағалау үшін ғылыми мәліметтердің ішкі жиынында сыналды. Өнімділікті бағалау үшін Precision, recall және F1-score көрсеткіштері пайдаланылды. Сонымен қатар, оның тиімділігін бағалау үшін мәтінді талдаудың басқа құралдарымен салыстырмалы талдау жүргізілді.

Модульдің жұмыс алгоритмі оның компоненттерінің дәйекті өзара әрекеттесуіне негізделген, бұл деректерді дәйекті өңдеуге, жіктеуге және нәтижелерді шығаруға мүмкіндік береді. Модульдің әртүрлі бөліктерінің қалай жұмыс істейтінін жақсы түсіну үшін төменде компоненттер арасындағы өзара әрекеттесу үшін код үзінділері берілген.

Мысал ретінде:

```
class_probs = {'science': 0.6, 'literature': 0.4}
word_probs = {
   'science': {'data': 0.3, 'analysis': 0.2, 'experiment': 0.1},
   'literature': {'book': 0.4, 'novel': 0.3, 'author': 0.2}
}
```

Модуль мәтінді файлдан немесе тікелей пәрмен жолы арқылы қабылдайды.

Мәтін тазартылады, таңбалауыштарға бөлінеді, сүзіледі және лемматизацияланады.

Мысалға "Мақаланың дереккөздері сай келмеді" мәтіні ['мақаланың', 'дереккөздері', 'сай', 'келмеді'] деп талданады.

Алгоритм келесідей жұмыс істейді:

1. Әр санат (A) үшін ықтималдық логарифмдері есептеледі:

$$\log P(A|B) = \log P(A) + \sum_{\omega \in B} \log P(\omega \in B)$$

Бұл оларды көбейту кезінде ықтималдықтың төмен мәніне қатысты мәселелерден аулақ болады.

2. Максималды $\log P(A|B)$ мәні бар санат жиынтық ретінде таңдалады.

Ғылым санаты үлкен мәнге ие, сондықтан мәтін "science" ретінде жіктеледі.

Талдау нәтижесі пайдаланушыға көрсетіледі.

Жақсартулар

Бастапқы тестілеу нәтижелеріне сүйене отырып, алгоритм жіктеу дәлдігін жақсарту және өңдеу уақытын қысқарту мақсатында жетілдірілді. Атап айтқанда, белгілерді таңдау процесі оңайландырылды және сәйкес және маңызды емес мәтіндерді жақсырақ ажырату үшін ықтималдық шектері түзетілді.

Талқылау

Ғылыми мәтіндерді талдау үшін Байес алгоритмін құру мәтіндерді жіктеудің тиімділігі мен дәлдігін арттыруда айтарлықтай әлеует көрсетті. Алгоритмнің алдын-ала ықтималдыққа негізделген белгілі бір терминдер мен сөз тіркестерінің өзектілігін болжау қабілеті оған дәстүрлі кілт сөздерді іздеу әдістерінен асып түсуге мүмкіндік берді. Мәтінмәндік өзектілік пен сөз жиілігін ескере отырып, алгоритм ғылыми мақалалардың мазмұны туралы толық түсінік алуға мүмкіндік береді.

[5] Мәтінді талдауға әсері. Әзірленген алгоритмнің негізгі артықшылықтарының бірі-оның масштабталуы және әртүрлі зерттеу салаларына бейімделуі. Ол ғылыми мәтіндердің үлкен көлемін адамның минималды араласуымен өңдей алады, бұл әсіресе жаңа зерттеулер тез жарияланатын қазіргі жағдайда өте маңызды. Сонымен қатар, алгоритм тестілеу кезінде дәлдік пен есте сақтау көрсеткіштерін жақсарта отырып, маңызды емес ақпаратты сүзудің жоғары қабілетін көрсетті. Бұл алгоритм тек әдеби шолулар үшін ғана емес, сонымен қатар жүйелі шолулар немесе мета-талдаулар сияқты арнайы қолданбалар үшін де пайдалы болуы мүмкін екенін көрсетеді.

Мәселелер мен шектеулер. Артықшылықтарына қарамастан, алгоритм бірқатар мәселелерге тап болды. Біріншіден, Байес ықтималдығына сүйену талдаудың дәлдігі оқу деректер жиынтығының сапасы мен өкілдігіне қатты тәуелді екенін білдіреді. Егер бастапқы деректер жиынтығы біржақты немесе әр түрлі болмаса, алгоритм бұрмаланған нәтижелер бере алады. Тағы бір шектеу-алгоритм құрылымдық және жартылай құрылымдалған мәтіндермен тиімді жұмыс істегенімен, гуманитарлық және әлеуметтік ғылымдар мәтіндерінде жиі кездесетін өте күрделі немесе екіұшты тұжырымдармен жұмыс істемейді. Бұл белгілерді бөлектеу процесін одан әрі нақтылауды қажет етуі мүмкін.

Сонымен қатар, алгоритм ғылыми дәлелдердің нюанстарын анықтауда қиындықтарға тап болды, мұнда белгілі бір тұжырымдар тікелей айтылғаннан гөрі көбірек болуы мүмкін. Бұл алгоритмнің болашақ итерациялары семантикалық талдау немесе үлкен және әртүрлі деректер жиынтығында оқытылған Машиналық оқыту үлгілері сияқты табиғи тілді (natural Language Understanding - NLU) түсінудің тереңірек мүмкіндіктерін қамтуы керек екенін көрсетеді.

Болашақ бағыттар. Әрі қарайғы жұмыс бағыттарының бірі-алгоритмге өз қателіктерінен сабақ алуға және уақыт өте келе өнімділігін үнемі жақсартуға мүмкіндік беретін Машиналық оқыту әдістерін енгізу. Трансформаторлар сияқты тереңірек NLP модельдерін біріктіру алгоритмнің күрделі тілдік құрылымдарды өңдеу қабілетін арттырып, терминдер арасындағы қатынастарды жақсырақ түсіндіре алады. Сонымен қатар, алгоритмді көп тілді

деректер жиынтығында тестілеу оның қолданылуын кеңейтіп, халықаралық зерттеулерді қолдауға мүмкіндік береді.

Тағы бір ықтимал бағыт - алгоритмнің мүмкіндіктерін кеңейту ғылыми мәтіндерді қорытындылау, бұл зерттеушілерге бүкіл құжатты оқымай-ақ мақаланың негізгі ойларын тез анықтауға көмектеседі. Бұл әсіресе медицина немесе технология сияқты қарқынды дамып келе жатқан салаларда жедел шолулар жүргізу кезінде пайдалы болуы мүмкін.

Қорытынды

Байес тәсіліне негізделген ғылыми мәтіндерді талдау алгоритмін әзірлеу әдебиеттерді талдаудың тиімділігі мен дәлдігін арттырудағы маңызды қадам болып табылады. Ықтималдық теориясын қолдана отырып, алгоритм зерттеушілерге жұмыс процесін оңтайландырудың қуатты құралын ұсына отырып, үлкен деректер массивтеріндегі сәйкес мазмұнды дұрыс жіктей және бағалай алады. Нәтижелер алгоритмнің құрылымдалған және жартылай құрылымдалған мәтіндермен жұмыс жасауда дәл және мағыналы тұжырымдар жасауда әсіресе тиімді екенін көрсетеді.

Дегенмен, зерттеу сонымен қатар бірқатар мәселелерді анықтады, әсіресе күрделі лингвистикалық құрылымдармен жұмыс істеуде және оқу деректер жиынтығының жеткілікті әртүрлілігін қамтамасыз етуде. Қазіргі алгоритм көптеген контексттерде жақсы жұмыс істегенімен, Машиналық оқыту модельдерін біріктіру және тілді тереңірек түсіну үшін оның мүмкіндіктерін кеңейту сияқты болашақ жақсартулар оның әмбебаптығы мен өнімділігін одан әрі арттырады.

Қорытындылай келе, бұл алгоритм ғылыми мәтіндерді талдау саласына перспективалы үлес қосатынын атап өткен жөн. Әрі қарай жетілдіре отырып, ол әртүрлі пәндердегі зерттеушілер үшін баға жетпес ресурс бола алады, бұл неғұрлым негізделген шешімдер қабылдауға және ғылыми жаңалықтардың қарқынын жеделдетуге мүмкіндік береді.

Пайданылған әдебиеттер

1. Bishop, C. M. «Pattern Recognition and Machine Learning. Springer», 2006.
2. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. «Bayesian Data Analysis. CRC Press», 2013.
3. Manning, C. D., Raghavan, P., & Schütze, H. «Introduction to Information Retrieval. Cambridge University Press», 2008.
4. Aggarwal, C. C., & Zhai, C. (Eds.). «Mining Text Data. Springer», 2012.
5. Blei, D. M., Ng, A. Y., & Jordan, M. I. «Latent Dirichlet Allocation. Journal of Machine Learning Research», 3, 993–1022, 2003.

ПОСТРОЕНИЕ АЛГОРИТМА И ЕГО ВЛИЯНИЕ НА АНАЛИЗ НАУЧНЫХ ТЕКСТОВ

Тұрарбек Әмина Мұратбекқызы

Научные руководители: Муханова А.А., PhD, Алтынбек С.А.

В данной статье рассматривается разработка алгоритма анализа научных текстов с применением Байесовского метода. В условиях стремительного роста числа научных публикаций эффективное извлечение информации приобретает особую значимость. В работе анализируются существующие методы обработки текстов, включая лексический анализ, машинное обучение и Байесовские модели, а также выявляются их преимущества и недостатки. Предлагаемый алгоритм использует Байесовскую вероятность для классификации и структурирования научных данных, обеспечивая высокую точность анализа. Алгоритм способен повысить эффективность обработки научных текстов, снизить количество ошибок и автоматизировать процесс их классификации. В перспективе планируется интеграция глубоких нейронных сетей и расширение возможностей модели для работы с многоязычными корпусами.

Ключевые слова: Байесовский метод, анализ научных текстов, машинное обучение, классификация данных, обработка текста, алгоритм, нейронные сети, автоматизация, лексический анализ, вероятностные модели.

ALGORITHM DEVELOPMENT AND ITS IMPACT ON SCIENTIFIC TEXT ANALYSIS

Turarbek A.M.

Scientific Supervisors: Mukhanova A.A., Altynbek S.A.

This article explores the development of a scientific text analysis algorithm using the Bayesian approach. Given the rapid increase in the number of scientific publications, efficient information retrieval has become crucial. The study examines existing text processing methods, including lexical analysis, machine learning, and Bayesian models, identifying their advantages and limitations. The proposed algorithm utilizes Bayesian probability to classify and structure scientific data, ensuring high analysis accuracy. The algorithm enhances the efficiency of scientific text processing, reduces errors, and automates the classification process. Future plans

include integrating deep neural networks and expanding the model's capabilities for multilingual corpora.

Keywords: Bayesian method, scientific text analysis, machine learning, data classification, text processing, algorithm, neural networks, automation, lexical analysis, probabilistic models.

REFERENCES

1. Bishop, C. M. «Pattern Recognition and Machine Learning. Springer», 2006.
2. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. «Bayesian Data Analysis. CRC Press», 2013.
3. Manning, C. D., Raghavan, P., & Schütze, H. «Introduction to Information Retrieval. Cambridge University Press», 2008.
4. Aggarwal, C. C., & Zhai, C. (Eds.). «Mining Text Data. Springer», 2012.
5. Blei, D. M., Ng, A. Y., & Jordan, M. I. «Latent Dirichlet Allocation. Journal of Machine Learning Research», 3, 993–1022, 2003.