https://doi.org/10.31891/2219-9365-2023-73-1-23
УДК 004.89

Роман ШАПТАЛА
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"
https://orcid.org/0000-0002-4367-5775
e-mail: r.shaptala@gmail.com
Геннадій КИСЕЛЬОВ
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"
https://orcid.org/0000-0003-2682-3593
e-mail: g.kyselov@gmail.com

МЕТОД ЗЛИТТЯ БАГАТОМОДАЛЬНИХ ВЕКТОРНИХ ПРЕДСТАВЛЕНЬ СЛІВ У МАЛОРЕСУРСНОМУ СЕРЕДОВИЩІ

У даній статті представлено метод злиття багатомодальних векторних представлень слів у малоресурсному середовищі. Цей метод, на відміну від інших методів злиття векторних представлень слів, враховує обмеження малоресурсного середовища і дозволяє поєднувати вектори слів з різних джерел, таких як документи та словники. Метод покладається на обчислення міжрядкової відстані замість побудови повних синтаксичних і морфологічних моделей, що часто неможливо в малоресурсних мовах. Його можна використовувати на проміжних етапах побудови систем обробки природної мови та машинного навчання при вирішенні практичних завдань, таких як машинний переклад чи класифікація документів.

Крім того, проведено аналіз різних методів злиття багатомодальних векторних представлень слів у малоресурсному середовищі. У статті описуються переваги, недоліки та обмеження кожного підходу, враховуючи завдання побудови уніфікованого векторного представлення тексту в поєднанні з даними з додаткових джерел. У дослідженні прикладом завдання у малоресурсному середовищі була обрана класифікація петицій до Київської міської ради, написаних українською мовою.

Велика кількість функцій обчислення міжрядкової відстані ускладнює їх вибір при вирішенні практичних задач. Ми пропонуємо набір рекомендацій у контексті малоресурсних середовищ, а також методологію вибору найкращого для вирішення поставлених завдань. Проаналізовані функції обчислення міжрядкової відстані включають відстань Левенштейна, подібність Жаккара, Мангеттенську відстань, відстань Хеммінга та коефіцієнт Дайса. Наші результати демонструють, що метод на основі відстані Левенштейна збільшує якість класифікації документів сильніше, ніж альтернативи. Ці висновки мають практичне значення для різних сфер, включаючи обробку природної мови, аналіз текстів та пошук інформації.

Ключові слова: *машинне навчання, обробка природної мови, математичне моделювання, нейронні мережі, векторні представлення слів, міжрядкова відстань.*

Roman SHAPTALA, Gennadiy KYSELOV
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

METHOD OF MULTIMODAL WORD EMBEDDINGS FUSION IN A LOW-RESOURCE SETTING

This paper presents a method of multimodal word embeddings fusion in a low-resource setting. This method, unlike other methods of word embeddings fusion, takes into account the limitations of a low-resource environment, and allows combining word embeddings from different sources, such as documents and dictionaries. The method relies on string distance calculations instead of building complete syntactic and morphological models which is often impossible in low-resource languages. This method can be used at intermediate stages of building natural language processing systems and machine learning when solving practical problems, such as machine translation or document classification.

Additionally, we present an analysis of various multimodal word embeddings fusion methods in a low-resource setting. The paper describes the advantages, disadvantages, and limitations of every approach given the task of building unified vector representation of text combined with data from additional sources. As an example of low-resource environment, our study shows the efficacy of described methods on the task of classifying Kyiv City petitions written in Ukrainian language.

Abundance of string distance functions makes the choice of one a difficult task. We propose a set of recommendations in the context of low-resource settings as well as a methodology to select the best one given practical task. Analyzed string distance functions include Levenstein distance, Jaccard similarity, Block distance, Hamming distance, and Dice's coefficient. Our results demonstrate that Levenshtein distance is more effective than others in this context, providing insights into which methods are best suited for low-resource analysis of string data. These findings have practical implications for various fields, including natural language processing, text mining, and information retrieval.

Keywords: *machine learning, natural language processing, mathematical modelling, neural networks, word embeddings, string distance.*

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Злиття векторних представлень слів є важливим етапом у багатьох завданнях машинного навчання, оскільки векторні представлення слів можуть використовуватись для розв'язання різноманітних задач, таких як класифікація текстів, машинний переклад, пошук інформації, та розпізнавання мови. Злиття векторних

представлень слів полягає у створенні єдиного вектора для кожного слова, який містить в собі інформацію про всі аспекти цього слова, що були зібрані з різних джерел. Це дозволяє зменшити розмірність даних, що зменшує витрати на обробку та збереження великої кількості векторів [1].

Крім того, злиття векторних представлень може покращити якість векторів за рахунок комбінації інформації з різних джерел, таких як різні мовні корпуси чи словники, через що такі представлення називаються багатомодальними. Це може допомогти вирішити проблему недостатньої кількості даних у малоресурсних середовищах та покращити якість відповідних моделей машинного навчання.

Якість методів злиття багатомодальних векторних представлень можна перевіряти кількома видами метрик, а саме внутрішніми - порівняння злитих представлень з оригінальними за відстанню у багатовимірному просторі, та зовнішніми - за впливом на результат повної моделі для вирішення конкретної задачі. У статті перевіряється якість на основі зовнішньої метрики, тому що, попри інтуїтивність внутрішніх метрик, злиття векторних представлень є виключно проміжним етапом вирішення більшого завдання.

Малоресурсність мовного середовища можна визначити як обмеженість ресурсів для навчання моделей машинного навчання при вирішенні задач обробки природної мови. Це можуть бути обмеження в обсязі даних, недостатність наборів даних для побудови якісних моделей, або їх недоступність через правові або технічні обмеження. Це явище особливо актуальне для менш розповсюджених та непопулярних у мережі Інтернет мов або діалектів, де обмежена доступність до достатньої кількості текстів для навчання моделей.

У таких середовищах, зокрема, може бути мінімальна кількість текстових даних, відсутність великих корпусів текстів або обмежена доступність мовних ресурсів. Це впливає на якість, а часто і унеможливлює створення, векторних представлень слів, які є ключовим елементом у багатьох завданнях машинного навчання, таких як класифікація тексту, пошук інформації, розпізнавання мови та інші [2].

Отже, розробка ефективних методів злиття багатомодальних векторних представлень слів у малоресурсному середовищі є актуальною задачею для покращення якості моделей машинного навчання та обробки природної мови в таких умовах.

Аналіз досліджень та публікацій

З все частішим використанням багатошарових нейронних мереж для вирішення задач обробки природної мови, дослідження у їх інтерпретації та аналізі проміжних шарів набули популярності. Через новизну та високу якість найбільш популярними векторними представленнями слів стали Word2Vec [3] та FastText [4]. Подібні векторні представлення кодують лише певний аспект слова, тому дослідження методів злиття векторних представлень слів, які кодують різні значення чи особливості мовної одиниці успішно проводилися у роботах [5] та [6]. Цікавим також є підхід авторів [7], які використовують модель суміші Гауса як метод злиття багатомодальних векторних представлень слів. У їх експериментах модель навчається на двох наборах даних, які сумарно вміщують 3.5 мільярди слів. Так як метод вимагає великої кількості даних для тренування, його користь у малоресурсному середовищі - обмежена. Аналогічну проблему досліджують у [8] - на основі 300 мільйонів речень автори будують багатомодальні вектори слів на основі рекурентних нейронних мереж. Такі моделі відомі високими вимогами до обчислювальних ресурсів та даних для ефективної роботи, а отже майже не застосовуються у малоресурсних середовищах.

Окрім малої кількості досліджень щодо роботи методів злиття багатомодальних векторних представлень слів, автори рідко описують процес пошуку відповідних слів у різних модальностях. Так, враховуючи різні форми слів у реальних текстових даних, відповідність між однаковими сутностями у різних джерелах не є простою перевіркою рівності кількох рядків. Отже, модифікація методів злиття векторних представлень слів етапом пошуку відповідника за одним з критеріїв міжрядкової відстані дозволяє більш системно описувати вирішення практичних задач обробки природної мови на основі векторних представлень з кількох джерел.

Формулювання цілей статті

Мета статті полягає у розробці та оцінці якості нового методу для злиття багатомодальних векторних представлень слів у малоресурсному середовищі. Такий метод повинен враховувати специфічні обмеження малоресурсних середовищ, а також спиратись на критерії міжрядкової відстані для автоматичного пошуку відповідників серед джерел текстової інформації. Відповідно до поставленої проблеми та обмежень у попередніх дослідженнях метою даної роботи є аналіз та порівняння різних методів злиття багатомодальних векторних представлень слів в малоресурсних середовищах, а саме опис переваг, недоліків та обмежень різних підходів для створення рекомендацій вибору найкращих методів поєднання закодованих слів.

Виклад основного матеріалу

Методи злиття багатомодальних векторних представлень слів поділяють на дві великі групи - на основі простих операцій та на основі нейронних мереж [9]. До першої групи відносять такі методи як

конкатенацію, усереднення, та зважене усереднення. Друга група виділяється великою кількістю архітектур, кожна з яких підлаштована під наскрізну задачу, яка вирішується.

Методи злиття багатомодальних векторних представлень слів на основі простих операцій мають наступні переваги: простота та ефективність реалізації - такі методи дуже прості в реалізації та не вимагають великих обчислювальних ресурсів; гнучкість - такі методи можуть бути використані з будь-якими видами векторних представлень, що дозволяє їх використання в різноманітних задачах; висока інтерпретованість - злиття на основі простих операцій дозволяє легко інтерпретувати результати та зрозуміти, які особливості кожної модальності впливають на кінцевий результат. У той же час до їх обмежень відносяться обмежена експресивність - злиття на основі простих операцій не завжди може передати всю інформацію з різних модальностей, що може призвести до втрати точності вирішення задачі; чутливість до величини та розміру векторних представлень - такі методи можуть бути чутливі до розміру та величини векторних представлень, що може призвести до невірних результатів при злитті.

Методи злиття багатомодальних векторних представлень слів на основі нейронних мереж досить поширені в задачах обробки природньої мови. Основна ідея полягає в тому, щоб об'єднати векторні представлення слів з різних джерел, таких як текстові корпуси та словники, у єдину багатовимірну просторову модель, яка може бути використана для подальшої обробки. Такі підходи навчаються наскрізно та потребують відповідні сигнали у наборах даних для тренування. Це є їх найбільшою перевагою, адже злиття відбувається автоматично через підбір параметрів відповідно до вирішуваної задачі. Проте, саме через це їх застосування у контексті малоресурсних середовищ дуже обмежене, адже якісних розмічених наборів даних для наскрізного тренування таких моделей немає. Додатковим недоліком методів на основі нейронних мереж є те, що вони маю низьку швидкодію, або вимагають спеціальні обчислювальні ресурси - прискорювачі, наприклад графічні процесори.

У випадку роботи з різними текстовими джерелами, як правило, створюють окремі словники з векторами слів під кожне джерело, а потім навчають їх наскрізно або окремо. При цьому різні словоформи одних і тих же слів отримують вектори різної якості, яка залежить від повноти використання даних слів у наборі даних для тренування. У такому випадку, для покращення якості цих методів пропонується модифікувати методи злиття багатомодальних векторних представлень слів на основі простих операцій додавши окремий етап пошуку відповідних слів з різних джерел. Найбільш точним способом реалізації такого етапу було б створення окремих синтаксичних та морфологічних моделей для повного співставлення слів з різних джерел у різних словоформах. Але у малоресурсних середовищах побудова таких моделей значно ускладнюється через брак даних для тренування, а моделі на основі правил зазвичай мають низьке покриття через неоднозначність у природних мовах [10]. Тому ми пропонуємо метод, що не вимагає навчання, а працює на основі функцій міжрядкової відстані.

Отже, запропонований метод злиття багатомодальних векторних представлень слів у малоресурсному середовищі для двох джерел текстових даних полягає у наступному: (1) маючи відповідні словники для кожного джерела, для словоформи, що подається на вхід методу, знаходиться найближче слово у даних словниках за критерієм функції міжрядкової відстані; (2) до знайдених слів знаходиться векторне представлення відповідно джерелу, з якого воно походить; (3) знайдені представлення зливаються за допомогою методів злиття на основі простих операцій, наприклад конкатенації. Схематично метод представлено на рис. 1.

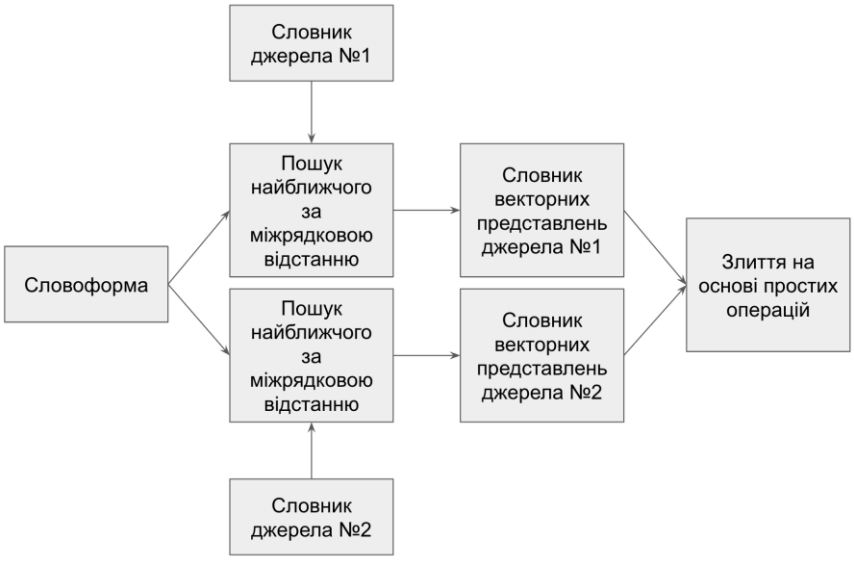

Рис. 1. Схема запропонованого методу

Для перевірки якості запропонованого методу та впливу вибору функції міжрядкової відстані було проведено експерименти у контексті малоресурсного середовища, а саме вирішення задачі класифікації петицій до Київської міської ради [11]. Набір даних містить 6560 петицій розмічених за 15 темами та був розділений на вибірки для тренування (80%) та тестування (20%). Другим джерелом текстової інформації виступав словник синонімів української мови, закодований за допомогою методів побудови графових представлень. Векторні представлення слів з документів побудовані за допомогою Word2Vec та представлення слів як вузлів графу словника синонімів подаються на вхід досліджуваним методам. Досліджувана метрика - F1-міра класифікації - є зовнішньою метрикою для методів злиття багатомодальних векторних представлень та застосовується у випадках класифікації з незбалансованим розподілом класів. Було проведено експерименти з багатошаровими нейронними мережами у якості методу злиття на основі нейронних мереж, які навчались на вищезазначеному наборі даних наскрізно, гіперпараметри та архітектури підбирались за допомогою методу сіткового пошуку. Серед методів простих операцій розглядались методи конкатенації, усереднення та зваженого усереднення, найкращим серед яких виявився метод зваженого усереднення. У запропонованому методів він також показав себе найкраще за F1-мірою. Результати порівняння якості запропонованого методу з методами інших класів наводяться у таблиці 1.

Таблиця 1

Вплив вибору методу злиття на якість класифікації у малоресурсному середовищі

Метод злиття	F1-міра
На основі нейронних мереж	0.606
На основі простих операцій	0.628
Запропонований метод	0.659

Ключовим елементом запропонованого методу є функція міжрядкової відстані, яка напряму впливає на якість злиття векторних представлень з різних джерел. Популярними функціями міжрядкової відстані є відстань Левенштейна [12], подібність Жаккара [13], Мангеттенська відстань [14], відстань Хемінга [15] та коефіцієнт Дайса [16]. Кількісні показники впливу вибору функції міжрядкової відстані у запропонованому методі на результат класифікації можна побачити у таблиці 2.

Таблиця 2

Вплив вибору функції міжрядкової відстані у запропонованому методі

Функція міжрядкової відстані	F1-міра
Відстань Левенштейна	0.659
Подібність Жаккара	0.625
Мангеттенська відстань	0.601
Відстань Хеммінга	0.645
Коефіцієнт Дайса	0.636

На основі вищезазначених експериментів було сформульовано набір рекомендацій щодо вибору найбільш ефективної функції міжрядкової відстані: (1) якщо обидва джерела текстових даних складаються з слів природної мови, варто використовувати відстань Левенштейна, адже вона враховує усі можливі варіанти редагування рядків; (2) можна застосовувати відстань Хеммінга у випадку якщо відстань Левенштейна є занадто обчислювально вимогливою для практичного рішення з мінімальною втратою якості; (3) інші функції міжрядкової відстані варто використовувати не з словами природної мови, а з послідовностями символів, такими як шифри чи коди.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

У даній статті було представлено новий метод злиття багатомодальних векторних представлень, що враховує специфічні обмеження малоресурсних середовищ, такі як відсутність великих наборів даних, і працює на основі обчислення міжрядкової відстані для пошуку відповідних слів з різних джерел. Експериментальні результати демонструють ефективність цього методу на прикладі класифікації петицій міста Києва, написаних українською мовою, забезпечуючи практичний та ефективний підхід для створення систем обробки природної мови та рішень машинного навчання.

Крім того, було проведено аналіз різних методів злиття багатомодальних векторних представлень в умовах обмежених ресурсів, та описано переваги, недоліки та обмеження різних підходів. Також пропонується набір рекомендацій щодо вибору найбільш ефективної функції міжрядкової відстані з огляду на практичне завдання класифікації документів. Для розглянутої задачі такою виявилась відстань Левенштейна.

Загалом, ця стаття робить внесок у сферу обробки природної мови, аналізу тексту та пошуку інформації. Беручи до уваги специфічні обмеження цих середовищ і пропонуючи більш практичний і

ефективний підхід, це дослідження має важливі наслідки для різних практичних застосувань, таких як машинний переклад і класифікація документів.

Хоча досягнуті результати демонструють ефективність запропонованого методу та вплив вибору функції міжрядкової відстані на якість моделі класифікації документів, у роботи є кілька обмежень, які залишають місце для подальших досліджень. По-перше, наше дослідження було зосереджено на конкретному малоресурсному середовищі (петиції до Київської міської ради, написані українською мовою), тому ефективність запропонованого нами методу може відрізнятися для інших мов і наборів даних. Таким чином, майбутні дослідження можуть додатково оцінити продуктивність цього методу у інших малоресурсних середовищах. По-друге, робота була зосереджена на задачі класифікації документів, але існує багато інших практичних застосувань багатомодального векторного представлення слів, таких як пошук інформації чи машинний переклад. Таким чином, майбутні дослідження можуть продовжити вивчення ефективності запропонованого методу та проаналізованих підходів у цих та інших практичних застосуваннях.

Література

1. Rouvier M., Favre B. SENSEI-LIF at SemEval-2016 Task 4: Polarity embedding fusion for robust sentiment analysis. SemEval@NAACL-HLT. 2016. P. 202-208.
2. Hedderich M.A., Lange L., Adel H., Strötgen J., Klakow D. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021. P. 2545-2568.
3. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.
4. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching word vectors with subword information. Transactions of the association for computational linguistics. 2017. P. 135-46.
5. Liang H., Lei W., Wang J., Jatowt A., Yang Z. From Unimodal to Multimodal Word Embeddings: A Survey. 2021.
6. Fukui K., Oshikiri T., Shimodaira H. Spectral graph-based method of multimodal word embedding. Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing. 2017. P. 39-44.
7. Athiwaratkun B., Wilson A. Multimodal Word Distributions. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. №1. P. 1645-1656.
8. Mao J., Xu J., Jing K., Yuille A.L. Training and evaluating multimodal word embeddings with large-scale web annotated images. Advances in neural information processing systems. №29. 2016.
9. Zhang C., Yang Z., He X., Deng L. Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing. 2020. 14(3). P. 478-93.
10. Wiechetek L., Pirinen F., Hämäläinen M., Argese C. Rules ruling neural networks–neural vs. rule-based grammar checking for a low resource language. Proceedings of the International Conference Recent Advances In Natural Language Processing 2021. 2021. P. 1530-1539.
11. Samvelyan, A., Shaptala, R., Kyselov, G. Exploratory data analysis of Kyiv city petitions. 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC). 2020. P. 1-4.
12. Левенштейн В. Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академии наук. 1965. №163(4). P. 845–848.
13. Jaccard, P. The distribution of the flora in the alpine zone. New Phytologist. 1912. P. 37-50.
14. Krause E. Taxicab Geometry: An Adventure in Non-Euclidean Geometry. Dover Books on Mathematics Series. 1986.
15. Hamming, R. Error detecting and error correcting codes. The Bell System Technical Journal. 1950. №29 (2). P. 147–160.
16. Dice L. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945. №26 (3). P. 297–302.

References

1. Rouvier M., Favre B. SENSEI-LIF at SemEval-2016 Task 4: Polarity embedding fusion for robust sentiment analysis. SemEval@NAACL-HLT. 2016. P. 202-208.
2. Hedderich M.A., Lange L., Adel H., Strötgen J., Klakow D. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021. P. 2545-2568.
3. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.
4. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching word vectors with subword information. Transactions of the association for computational linguistics. 2017. P. 135-46.
5. Liang H., Lei W., Wang J., Jatowt A., Yang Z. From Unimodal to Multimodal Word Embeddings: A Survey. 2021.

6. Fukui K., Oshikiri T., Shimodaira H. Spectral graph-based method of multimodal word embedding. Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing. 2017. P. 39-44.
7. Athiwaratkun B., Wilson A. Multimodal Word Distributions. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. №1. P. 1645-1656.
8. Mao J., Xu J., Jing K., Yuille A.L. Training and evaluating multimodal word embeddings with large-scale web annotated images. Advances in neural information processing systems. №29. 2016.
9. Zhang C., Yang Z., He X., Deng L. Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing. 2020. 14(3). P. 478-93.
10. Wiechetek L., Pirinen F., Hämäläinen M., Argese C. Rules ruling neural networks–neural vs. rule-based grammar checking for a low resource language. Proceedings of the International Conference Recent Advances In Natural Language Processing 2021. 2021. P. 1530-1539.
11. Samvelyan, A., Shaptala, R., Kyselov, G. Exploratory data analysis of Kyiv city petitions. 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC). 2020. P. 1-4.
12. Levenshtejn V. Dvoichnye kody s ispravleniem vypadenij, vstavok i zameshhenij simvolov. Doklady Akademii nauk. 1965. №163(4). P. 845–848.
13. Jaccard, P. The distribution of the flora in the alpine zone. New Phytologist. 1912. P. 37-50.
14. Krause E. Taxicab Geometry: An Adventure in Non-Euclidean Geometry. Dover Books on Mathematics Series. 1986.
15. Hamming, R. Error detecting and error correcting codes. The Bell System Technical Journal. 1950. №29 (2). P. 147–160.
16. Dice L. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945. №26 (3). P. 297–302.