УДК 512.563:81’374

Застосування алгебричних структур у комп’ютерній лексикографії

Богдан Філь[1], Ігор Кульчицький[2]

[1] к. ф.-м. н., доцент, Національний університет «Львівська політехніка», вул. С. Бандери, 12, Львів, 79013, e-mail: <post.me.now@gmail.com>
[2] к. т. н., доцент, Національний університет «Львівська політехніка», вул. С. Бандери, 12, Львів, 79013, e-mail: <bis.kim@gmail.com>

У статті зроблено спробу застосувати математичні методи до теорії побудови лексикографічних систем, зокрема до формалізації побудови словників. Під формалізованим поняттям словника розуміємо абстрактний мовно-інформаційний об’єкт, визначальною рисою якого, передусім, є членоване розміщення матеріалу. Основною композиційною та комунікативною одиницею слугує відносно самостійний відрізок тексту, який називають словниковою статтею. Описаний підхід доводить можливість застосування засобів і методів класичної алгебри до словникових структур. У праці введено адитивну та мультиплікативну операції над словниками та перевірено аксіоматику теорії множин (аксіоми об’ємності, суми, різниці та існування). Також доведено основні закони алгебри, які дають змогу стверджувати, що множину словників можна розглядати як класичну алгебру. Проведене дослідження властивостей множини словників показує, що можливо застосовувати до дослідження творення словників та їх аналізу методи та результати теорії множин і класичної алгебри.

Ключові слова: лексикографія, словник, лінгвотехнологія, алгебра, теорія множин.

Вступ. У час бурхливого розвитку інформаційного суспільства актуалізувалася проблема застосування адекватних формальних моделей у лінгвістиці, зокрема у лексикографії. Саме вони уможливлюють створення сучасних інтелектуальних лінгвотехнологій у природомовних людино-машинних системах [1].

Під словником розумітимемо абстрактний мовно-інформаційний об’єкт, визначальною рисою якого передусім є членоване розміщення матеріалу — основною композиційною та комунікативною одиницею слугує відносно самостійний відрізок тексту, який називають словниковою статтею [2]. Сукупність статей творить його основу. Врешті-решт інші елементи словника (передмова, джерельна база тощо) можна розглядати як статті виродженої структури, тому у сукупності словник можна розглядати як множину статей. У комп’ютерних технологіях роботу над словником розпочинають зі створення його абстрактного прототипу — порожнього словника. У процесі роботи до нього додають, вилучають статті та редагують їх.

Для того, щоб застосувати до дослідження словникотворення математичні методи, потрібно дати означення словника, яке б було позбавлене смислового

значення, і дало б можливість з аксіом, властивих алгебричним структурам, дедуктивним методом отримувати теореми. Хоча аксіоми ґрунтуються на інтуїтивному розумінні поняття словника, завдяки аксіоматичному методу інтуїтивне розуміння змісту цього поняття не буде використовуватися ні під час доведення теорем, ні в означеннях.

Позаяк абстрактна математика має справу з однорідними об'єктами, природу яких ігнорують, то, для однорідності, статтю, яку додають до словника чи вилучають з нього, можна розглядати як словник, що містить одну статтю. Одночасно, словникова стаття — елемент словника. Таке тлумачення словника дозволяє впровадити для опису словникових об'єктів поняття, властиві математичним об'єктам — операції над об'єктами.

Надалі ми можемо використовувати лексикографічні й алгебричні поняття, які є рівноцінними в межах нашого дослідження — стаття й елемент, словник і множина.

1. Алгебрична аксіоматика множин словників

Первинні поняття теорії множин — множина та відношення бути елементом множини. Замість x є множина будемо писати X, замість x є елементом множини Y — $x \in Y$. Заперечення виразу $x \in Y$ будемо записувати у вигляді $x \notin Y$ або $\neg(x \in Y)$. Для зручності, використовуватимемо великі латинські букви для позначення словників (множин), малі латинські букви для позначення словникових статей (елементів).

Використаємо чотири аксіоми з теорії множин:

1. *аксіома об'ємності*: якщо словники A та B складаються з тих самих статей, то вони співпадають;
2. *аксіома суми*: для довільних словників A та B існує словник, статті якого є всі статті словника A та всі статті словника B і який ніяких інших статей не містить;
3. *аксіома різниці*: для довільних словників A та B існує словник, статтями якого є ті і тільки ті статті словника A, які не є статтями словника B;
4. *аксіома існування*: існує хоча б один словник.

З 1-ої та 2-ої аксіом отримуємо наслідок, що словник, який є результатом застосування *аксіоми суми*, єдиний. Насправді, якщо б було два таких словники C_1 і C_2, які є сумою словників A та B, то вони обидва містили б ті самі статті (всі статті словника A та всі статті словника B) і тому згідно *аксіоми об'ємності* була б рівність $C_1 = C_2$. Цей словник (єдиний), який задовольняє *аксіому суми*, будемо називати сумою (або об'єднанням) словників A та B і позначатимемо символом $A \cup B$. Для довільного x та довільних словників A та B справедлива еквівалентність

$$x \in A \cup B \equiv (x \in A) \vee (x \in B).$$

Так само, з *аксіоми 1* та *аксіоми 3*, висновуємо, що для довільних словників A та B існує тільки один словник, якому належать статті словника A, що не належать словнику B. Цей словник називається різницею словників A та B і

позначається символом $A - B$. Для довільного x і довільних словників A та B справедлива еквівалентність

$$x \in A - B \equiv (x \in A) \wedge (x \notin B).$$

Із законів логіки (де Моргана та подвійного заперечення) також висновуємо, що

$$\neg(x \in A \cup B) \equiv \neg(x \in A) \vee (x \in B). \quad (1)$$

Тобто x не належить до різниці $A - B$, якщо x не належить до A або належить до B. За допомогою операцій «$\cup$» та «–» можна задати ще дві операції на словниках — добуток (перетин) і симетричну різницю.

Добуток (*перетин*) $A \cap B$ словників A та B визначаємо формулою

$$A \cap B = A - (A - B).$$

З означення різниці легко отримати для довільного x:

$$x \in A \cap B \equiv (x \in A) \wedge \neg(x \in A - B). \quad (2)$$

Зі співвідношення (2) за використання формули (1) і першого закону дистрибутивності у логіці остаточно отримуємо:

$$x \in A \cap B \equiv (x \in A) \wedge (x \in B).$$

Іншими словами *добуток словників* — це спільна частина співмножників. Статтями добутку є ті і тільки ті об'єкти, які належать до обох співмножників.

Словник A називається підсловником словника B, якщо кожна стаття словника A належать до словника B, тобто словник A міститься у словникові B. Таке відношення позначатимемо $A \subset B$ або $B \supset A$ і будемо називати *відношенням включення*. Відношення включення є транзитивне (це неважко показати):

$$(A \subset B) \wedge (B \subset C) \to (A \subset C).$$

Відношення включення можна визначити за допомогою відношення рівності та однієї з операцій додавання (об'єднання) $\cup$ або множення (перетину) $\cap$:

$$(A \subset B) \equiv (A \cap B = B) \equiv (A \cap B = A).$$

З *аксіоми 3* висновуємо — якщо існує хоча б один словник A, то існує словник $A - A$, який не містить жодної статті. Такий словник єдиний. Єдиний словник, який не містить жодної статті, називається *порожнім словником* і позначатимемо його символом 0. Рівність $A \cap B = 0$ означає, що словники A та B не мають спільних статей, або, інакше, — не перетинаються.

Рівність $A - B = 0$ означає, що $B \subset A$. Роль порожнього словника в теорії словників аналогічна ролі числа нуль в алгебрі. Якщо б не було словника 0 операції додавання та віднімання словників не завжди б можна було виконати, що породжувало б труднощі в подальшому під час дослідження властивостей словників.

2. Деякі закони теорії алгебр множин словників

Операції додавання, множення та віднімання словників мають багато спільного з відповідними операціями над числами. Вкажемо закони поведінки операцій, які відображають відповідні властивості, і ті, які не мають аналогій в арифметиці.

Закони комутативності

$$A \cup B = B \cup A, \quad A \cap B = B \cap A.$$

Ці закони безпосередньо висновуються із законів комутативності для диз'юнкції та кон'юнкції.

Закони асоціативності

$$A \cup (B \cup C) = (A \cup B) \cup C, \quad A \cap (B \cap C) = (A \cap B) \cap C.$$

Доведення базується на законах асоціативності для диз'юнкції та кон'юнкції.

Ці закони означають, що сума (добуток) скінченого числа словників не залежить ні від порядку виконання операцій, ні від способу групування операцій, тобто чи сумуємо (множимо) спочатку окремі доданки (співмножники), чи розбиваємо на окремі групи, а потім додаємо (перемножуємо) результати операцій на групах, отримаємо той самий результат.

Закони дистрибутивності:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

Доведення ґрунтується на законах дистрибутивності кон'юнкції щодо диз'юнкції та диз'юнкції стосовно кон'юнкції. Перший закон дистрибутивності повністю аналогічний закону дистрибутивності в арифметиці, якщо замінити знак $\cap$ на знак множення (*), та знак $\cup$ на знак додавання (+). Для другого ж закону дистрибутивності не має відповідного в арифметиці.

Закони ідемпотентності

$$A \cup A = A, \qquad A \cap A = A.$$

Доведення слідує з відповідних законів математичної логіки.

Використовуючи закони дистрибутивності, доводимо закон

$$A \cup (B - A) = A \cup B.$$

З цього закону висновується, що віднімання словників не є операція, обернена до додавання, тобто $A \cup (B - A) \neq B$, хоча, якщо $A \subset B$, то ця нерівність перетворюється в рівність, як і у звичайній арифметиці.

Подамо, без доведення, ще кілька законів для різниці словників

$$A - B = A - (A \cap B),$$

закон дистрибутивності множення стосовно віднімання

$$A \cap (B - C) = (A \cap B) - C.$$

Закони де Моргана в алгебрі словників мають вигляд

$$A-(B\cap C)=(A-B)\cup(A-C),\quad A-(B\cup C)=(A-B)\cap(A-C).$$

Доведення цих законів базується на законах де Моргана алгебри висловлювань.

Випишемо, без доведення, ще кілька рівностей

$$(A\cup B)-C=(A-C)\cup(B-C),\quad A-(B-C)=(A-B)\cup(A\cap C),$$
$$A-(B\cup C)=(A-B)-C.$$

Подальші імплікації ілюструють аналогію між відношеннями включення та відношенням «не більше» в арифметиці:

$$(A\subset B)\wedge(C\subset D)\to\left[(A\cup C)\subset(B\cup D)\right],$$
$$(A\subset B)\wedge(C\subset D)\to\left[(A\cap C)\subset(B\cap D)\right],$$
$$(A\subset B)\wedge(C\subset D)\to\left[(A-D)\subset(B-C)\right],$$
$$(C\subset D)\to\left[(A-D)\subset(A-C)\right].$$

Додамо ще означення симетричної різниці словників.

Симетричну різницю $A\div B$ двох словників *A* та *B* визначимо як

$$A\div B\equiv(A-B)\cup(B-A).$$

До симетричної різниці належать статті, що належать до словника *A*, але не належать до *B*, та статті, що належать до словника *B*, але не належать до *A*.

Ця операція має властивості комутативності й асоціативності:

$$A\div B=B\div A,\quad A\div(B\div C)=(A\div B)\div C.$$

Операція добутку дистрибутивна щодо симетричної різниці:

$$A\cap(B\div C)=(A\cap B)\div(A\cap C).$$

Порожній словник є нульовий елемент: $A\div A=0$.

Отож, якщо, до цього не було різниці між операціями симетричної різниці та додавання, то тепер вона одразу кидається у вічі. Окрім цього, операція має обернену, тобто для довільних словників *A* та *C* існує і тільки один словник *B*, такий, що $A\div B=C$, а саме, $B=A\div C$. Нескладно довести, що $A\div(A\div C)=C$. Тобто операція симетричної різниці «÷» має обернену та ця операція є тією ж симетричною різницею.

Подані властивості показують, що словники утворюють алгебраїчне кільце, якщо «сумою» буде операція «÷», а «добутком» — операція «$\cap$». Цей факт важливий ще й тому, що операції «$\cup$» та «–» можна виразити через симетричну різницю та додавання:

$$A \bigcup B = A \div B \div \left(A \bigcap B \right), \quad A \div B = A \div \left(A \bigcap B \right).$$

Наслідком двох останніх формул є твердження: якщо два словники A та B не перетинаються, то $A \bigcup B = A \div B$.

Висновки. Описані властивості множини словників показують, що можливо застосовувати до дослідження творення словників та їх аналізу методи й результати теорії множин і класичної алгебри. А це дає можливість використовувати математичний апарат там, де традиційно використовувалися методики, притаманні гуманітарним наукам. Вводячи структуру алгебри на множині словників, відкриваємо шлях до використання не тільки характеристик, що притаманні наборам множин, але і для встановлення взаємних відповідностей між словниками. Таким чином, закладається основа для дослідження пов'язаності словників, наприклад дослідження відповідностей словників різних мов і, зокрема, алгебраїзації побудови перекладацьких словників.

Література

[1] *Широков В. А.* Феноменологія лексикографічних систем; НАН України. Укр. мов.-інформ. фонд. — Київ: Наук. думка, 2004. — 327 с.

[2] *Широков В. А.* Комп'ютерна лексикографія; НАН України. Укр. мов.-інформ. фонд. — Київ: Наук. думка, 2011. — 351 с.

[3] *Куратовский К., Мостовский А.* Теория множеств. — Москва: Мир, 1970. — 409 с.

[4] *Широков В. А.* Інформаційна теорія лексикографічних систем. — Київ: Довіра, 1998. — 331 с.

[5] *Бурбаки Н.* Теория множеств. — Москва: Мир, 1965. — 455 с.

Application of algebraic structures in computer lexicography

Bohdan Fil, Ihor Kulchytskyy

The article attempts to apply mathematical methods to the theory of creation of lexicographical systems, in particular to the formalization of dictionaries creation. The formalized concept of a dictionary we understand as an abstract language and information object, the key feature of which is, primarily, segmental placement of material. A relatively independent piece of text serves as the main compositional and communicative unit which is called a dictionary entry. The described approach demonstrates the possibility of applying tools and methods of classical algebra to the lexicographic structures. In this article, the additive and multiplicative operations on dictionaries were introduced, and axiomatic of the set theory (axioms of extensionality, sum, difference and existence) was checked. The basic laws of algebra, which allow to state that the set of dictionaries can be considered as classical algebra, were also proved. The conducted research of characteristics of dictionaries sets shows that methods and results of the set theory and classical algebra can be applied to the study of dictionaries creation and their analysis.

Применение алгебраических структур в компьютерной лексикографии

Богдан Филь, Игорь Кульчицкий

В статье сделана попытка применить математические методы теории построения лексикографических систем, в частности к формализации построения словарей. Под формализованным понятием словаря понимаем абстрактный языково-информационный объект, определяющей чертой которого, прежде всего, является расчлененное размещение материала. Основной композиционной и коммуникативной единицей служит относительно самостоятельный отрезок текста, который называют словарной статьей. Описанный подход доказывает возможность применения средств и методов классической алгебры в словарных структур. В работе введены аддитивная и мультипликативная операции над словарями и проверена аксиоматика теории множеств (аксиомы объемности, суммы, разности и существования). Также доказаны основные законы алгебры, которые позволяют утверждать, что множество словарей можно рассматривать как классическую алгебру. Проведенное исследование свойств множества словарей показывает, что можно применять к исследованию создания словарей и их анализа методы и результаты теории множеств и классической алгебры.

Представлено доктором технічних наук Я. П'янилом

Отримано 18.03.14