ИЖТИМОИЙ - СИЁСИЙ ТЕРМИНЛАР ТАРКИБИДАГИ СЎЗЛАРНИНГ ЧАСТОТАСИ

ҒИЁСОВ БОБУР
ўқитувчи, ТДШУ

Аннотация. *Мазкур мақола хитой тили ижтимоий-сиёсий матнларидан ижтимоий-сиёсий терминларни ажратиб олиш алгоритми ишлаб чиқилиб, ижтимоий-сиёсий терминлар таркибидаги сўзларнинг частотасини аниқлашни баён этади. Хитой тилидаги ижтимоий-сиёсий терминларни частота жадвали тузилиб, эксперимент ва компонент таҳлил методидан фойдаланиб, таҳлил этилган ва ижтимоий-сиёсий терминалар жадваллари тузилган. Мақолада битта сўздан ёки матнлар корпусидан бир ёки икки полисиллабик морфем иероглиф сўзларни автоматик равишда ажратиб олиш усуллари тасвирланган. Сўз бирикмаларининг "терминологиясини" ҳисоблашнинг беш хил вариантлари кўриб чиқилган. Тажрибалар турли хил билим соҳаларига оид учта маълумотлар тўпламларида ўтказилди. Қўшима баҳолаш техникаси таклиф этилади, методларни қиёсий баҳолаш натижалари тақдим этилади. Ишлаб чиқилган тизим доирасида натижалардан мумкин бўлган амалий фойдаланишнинг мисоли сифатида тегишли жадваллар билан частоталар ҳисоблаш усуллари ёритилди.*

Таянч сўз ва иборалар: *Хитой тилида ижтимоий-сиёсий терминлар, частота, стандартлаш, экспериментал метод.*

Аннотация. *В данной статье описывается разработка алгоритма выделения социально-политических терминов из китайских социально-политических текстов и определения частотности слов в социально-политических терминах. Были составлены частотные таблицы социально-политических терминов на китайском языке, проанализированы с использованием методов экспериментального и компонентного анализа, а также составлены таблицы социально-политических терминов. В статье описаны методы автоматического извлечения одного или двух многосложных морфемных иероглифических слов из одного слова или основного текста. Было рассмотрено пять различных вариантов расчета «терминологии» словосочетаний. Эксперименты проводились на трех наборах данных, относящихся к разным областям знаний. Предложены методики совместной оценки и представлены результаты сравнительной оценки методов. В качестве примера возможного практического использования результатов в рамках разработанной системы описаны методы расчета частот с соответствующими таблицами.*

Опорные слова и выражения: *Социально-политические термины на китайском языке, частота, стандартизация, экспериментальный метод.*

Abstract. *This article describes the development of an algorithm for extracting socio-political terms from Chinese socio-political texts and determining the frequency of words in socio-political terms. Frequency tables of socio-political terms in Chinese were compiled, analyzed using methods of experimental and component analysis, and tables of socio-political terms were compiled. The article describes methods for automatically extracting one or two polysyllabic morphemic hieroglyphic words from one word or the main text. Five different options for calculating the "terminology" of phrases were considered. The experiments were carried out on three datasets from different fields of knowledge. The methods of joint assessment are proposed and the results of the comparative assessment of the methods are presented. As an example of a possible practical use of the results within the framework of the developed system, methods for calculating frequencies with the corresponding tables are described.*

Keywords and expressions: *Socio-political terms in Chinese, frequency, standardization, experimental method.*

Кириш. Терминология тизимида ҳар бир сўзнинг мутлақ частотаси бир хил эмас. Баъзи сўзлар тез-тез ишлатилади ва улар юқори частотали сўзлар деб номланади ва баъзи сўзлар кўпинча нисбатан кам қўлланилиб, улар паст частотали сўзлар деб аталади. Терминларнинг номинатсия этишда, терминлар сони кўпайиши билан, юқори частотали сўзлар сони, қоида тариқасида, кўпаяди.

Глобаллашув жараёнида янги терминлар сони кўпайиши билан, турли тиллардаги терминларни маъноси жиҳатдан систем-стандартлашув масаласи талаб этилади, чунки бу ҳодиса фан ривожига жуда катта ҳисса қўшиб, турли тил чегара тўсиқларини келтириб чиқаради.Бунда ижтимоий-сиёсий матнлардаги терминларни системлаштириб, ижтимоий-сиёсий термин шакллланишида бир-бири билан бириккан сўзларни умумий частоталарни ҳисоблаймиз.

Ҳисоблаш усуллари. Хитой тилидаги сиёсий матнлардан ижтимоий - сиёсий терминлар ажратиб олишда муҳим вазифаларидан бири бу сиёсий матнларни маълум даражадаги ишончлилиги билан ҳужжатнинг тематик йўналишини акс эттирадиган терминларни ажратиб кўрсатишдир. Автоматик калит сўзларни матндан чиқариб олиш, ҳужжатдаги муҳим тематик терминларни - автоматик ажратиб кўрсатиш деб таърифланиши мумкин.

在术语系统中，每个单词的绝对頻率并不是一样的。有的单词经常使用，叫做高频词，有的单词不常使用，叫做低频词。

随着术语条目的增加，高频词的数目一般来说也相应地增加，而新闻出现的可能性越来越小。这时，尽管术语的条数还继续增加，不同单词总数增加的速率却越来越小，而高频词则反复地出现。

社会政治术语和中国政治文本的重要任务之一是将这些政治文本与能够在一定程度上反映文档主题的术语相区别。 自动关键词提取可以描述为文档中重要主题词的自动突出显示。[1]

Сўнгги йилларда ҳар хил ўлчамдаги ҳужжатлар тўпламини таҳлил қилиш ва бир ёки ундан ортиқ сўзлардан иборат ижтимоий - сиёсий терминларни ажратиб олиш имконини берадиган кўплаб ёндашувлар ишлаб чиқилмоқда.

Маълум бир ҳужжатдаги терминнинг маъносини аниқлаш учун аввал унга тегишли бўлган барча ҳужжатлар тўпламини таҳлил қилиш керак.

Таҳлилдан сўнг, илмий муқобил ёндашув усули орқали, маълум бир тилда мавжуд бўлган сўзлар тўпламининг озми-кўпми тахминий моделлари бўлган лингвистик онтологиялардан фойдаланиб, терминлар тизимлари яратилади.

Бу йўналишда натижаларнинг аниқлиги ва тўлиқлиги сабабли, шунингдек ажратиб олиш усулларидан фойдаланилиб, сиёсий матнлар, луғатлар, Интернет-қидирув тизимлари ёрдамида битта сўздан ёки матнлар корпусидан икки сўзли терминларни автоматик равишда ажратиб олиш имконини беради.

Ижтимоий - сиёсий терминлар таркибидаги сўзларнинг частоталари ҳисоблашнинг икки хил усули кўрилиб чиқилади.

Бу икки хил усул ижтимоий-сиёсий матнларда, маълумотлар тўпламларида ижтимоий - сиёсий терминлар таркибидаги сўзларнинг частоталари ҳисоблаш ўтказилади.

Қўшма баҳолаш техникаси таклиф этилади, методларни қиёсий баҳолаш натижалари тақдим этилади.

近年来，已经开发出许多方法，使您可以分析一组大小不同的文档，以及由一个或多个单词组成的单独的社会政治术语。要确定特定文档中术语的含义，您首先需要分析与之相关的整套文档。[2]

---

[1] 蔡梅 [Cai Mei], 现代汉语外来词的新形式及其规范问题，人才培养模式改 革和开放教育试点论文集，[xiàndài hànyǔ wàilái cí de xīn xíngshì jí qí guīfàn wèntí]. [New Forms of Modern Chinese Loanwords and Their Standardization Issues]. 北京，2003 年. 86-87 页 [The Commercial Press Beijing], 2003. pp. 86-87.

[2] 郑述谱 Zheng Shupu, 俄罗斯当代术语学 èluósī dāngdài shùyǔ xué [Russian Contemporary Terminology], 商务印书馆 shāngwù yìn shūguǎn [The Commercial Press]，2005 年. 24-27 页 pp. 24-27.

分析之后，将使用语言本体创建术语系统，该术语本体是使用替代科学方法以特定语言提供的一组单词的近似模型。

由于此方向上结果的准确性和完整性，以及使用分隔方法，因此可以使用政治文本，词典，网络搜索引擎从一个单词或主要文本中自动区分两个单词的术语。[1]

有两种从社会政治角度计算词频的方法。这两种方法用于计算社会政治文本中的词频，即社会政治术语中的数据集。

提出了联合评估的方法，并给出了方法比较评估的结果。[2]

Теримларнинг частотасини ҳисоблашнинг энг кенг тарқалган схемалари ТФ-ИДФ ва унинг турли хил вариантлари, шунингдек бошқалар (АТС, Окапи, ЛТУ).

Бироқ, ушбу схемаларнинг умумий хусусияти шундаки, улар барча ҳужжатлар тўпламидан маълумотларни талаб қилади.

Агар ҳужжат ваколатхонасини яратиш учун ТФ-ИДФ асосидаги усул ишлатилган бўлса, унда янги ҳужжатнинг тўпламга келиши барча ҳужжатлардаги терминлар частотасини қайта ҳисоблашни талаб қилади.

Шунинг учун, ҳужжатдаги терминлар частотасини қийматларига асосланган ҳар қандай дастурларга ҳам таъсир кўрсатилади.

Бу динамик маълумот оқимлари реал вақт режимида қайта ишланиши керак бўлган тизимларда ўқитишни талаб қиладиган асосий муддатли экстрактсия усулларидан фойдаланишни сезиларли даражада истисно қилади.

Ижтимоий-сиёсий терминларни шакллантиришда сўзларни бир зумда татбиқ этишга ва хорижий тиллардан хитой тилига сўзларни татбиқ этишда частотали луғатлар жуда керак.

T терминлари сони ва ***W*** ҳар хил сўзларнинг умумий сони ўртасида функтсионал боғлиқлик, *(1-жадвал мисолида)*

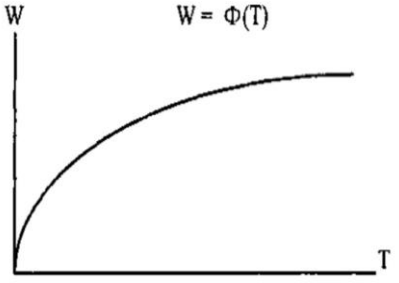
(在术语数 ***T*** *与不同单词总数* ***W*** *之间，存在着如下的函数关系):*

1- жадвал

$W = \phi\ (T)$

单词的术语构成频率就是在一个术语系统中运行单词的总数 P 被不同单词数 W 来除所得商。单词的术语构成频度用 Φ 表示。这样，可有下面的公式：

2- жадвал



[1] 冯志伟著 Feng Zhiwei，现代术语学引论 xiàndài shùyǔ xué yǐn lùn, [Introduction to Modern Terminology]，语文出版社，1997 年, 56-58 页 Language Publishing House, 1997, pages 56-58

[2]马菊红 Ma Juhong，科技术语翻译研究 kējì shùyǔ fānyì yánjiū, [Research on Translation of Technical Terms], 硕士论文，哈尔滨工业大学外国语言 学与应用语言学 shuòshì lùnwén, hā'ěrbīn gōngyè dàxué wàiguó yǔyán xué yǔ yìngyòng yǔyán xué ，[Harbin Institute of Technology Foreign Linguistics and Applied Linguistics] 1999 年. 46 页 1999. 46 pages

3- жадвал

$$\Phi = P / W$$

***T / W* = 1510 / 858 = 1.76，**

***T* = 1510** бўлганда, **ГЛОТ - С** тизимининг иқтисодий кўрсаткичи **1.76** ни ташкил қилади, ҳар бир сўзда ўртача **1.76** бирлик терминлар мавжуд, шунинг учун ушбу тизим юқори иқтисодий самарадорликка эга. Терминологик тизимнинг иқтисодий кўрсаткичи тизимдаги терминлар сонига боғлиқ. Иқтисодий индекснинг ўсиш суръати тизимдаги аъзолар сонининг кўпайиши билан янада юқори бўлиб бормоқда.

ГЛОТ - С да, тизимдаги терминлар сони **500** га ва турли хил сўзларнинг йиғиндиси **342** га тенг бўлса, унинг иқтисодий кўрсаткичи **1.46**, тизимдаги терминлар сони **1000** га кўпайтирилганда ва турли хил сўзларнинг йиғиндиси **588**га кўтарилганда, унинг иқтисодий кўрсаткичи айланади. **1.70** ва тизимдаги терминлар сони **1510** га кўпайганда ва турли хил сўзларнинг йиғиндиси **858** га кўпайганда, унинг иқтисодий кўрсаткичи **1.76** га тенг бўлади.

在这种情况下， $W = \sum \mathcal{V} = 588$

并且 $P = \sum \rho = 2072$

因此 $\Phi = P / W = 2072 / 588 = 3.52$

当系统中的术语数为 1000 条（T = 1510）时，单词的频率表如下：

4- жадвал

T	W	R	F
500	342	987	2.89
1000	588	2072	3.52
1510	858	3216	3.75

5- жадвал

$$\Phi = P / W = 3216/858 = 3.75$$

在这种情况下， $W = \sum \mathcal{V} = 858$

并且 $P = \sum \rho = 3216$

因此 $\Phi = P / W = 3216 / 858 = 3.75$

从面可得到如下的表：

6- жадвал

α	υ	ρ	α	υ	ρ	α	υ	ρ
1	411	411	13	5	65	26	2	52
2	150	300	14	3	42	27	2	54
3	73	219	15	2	30	33	2	66
4	52	208	16	3	48	34	1	34
5	44	220	18	1	18	35	1	35
6	24	144	19	3	57	44	1	44
7	14	98	20	1	20	47	1	47
8	14	112	21	2	42	55	1	55
9	13	117	22	1	22	56	1	56
10	5	50	23	2	46	63	1	63
11	8	88	24	3	72	68	1	68
12	7	84	25	2	50	79	1	79

这表明，当ГЛОТ-С系统中的社会政治用语数量为一千五百十时，其单词的周期性结构频率为三点七十五，即每个单词平均出现三点七十五次，因此该值也可以表示单词的内容。

这表明，当 ГЛОТ-С 系统中的社会政治用语数量为一千五百十时，其单词的周期性结构频率为三点七十五，即每个单词平均出现三点七十五次，因此该值也可以表示单词的内容。这些字。 单词结构的频率也会影响术语系统中的术语数量。

Бу шуни кўрсатадики, **ГЛОТ-С** тизимидаги ижтимоий-сиёсий терминлар сони **1510** га тенг бўлганда, унинг сўзларининг муддатли таркибий частотаси **3.75** ни ташкил қилади, яъни ҳар бир сўз ўртача **3.75** марта пайдо бўлиши мумкин, шунинг учун бу қиймат ушбу сўзларнинг таркибини ҳам англатиши мумкин.[1]

Сўзларнинг терминлар таркиби частотаси терминологик тизимдаги терминлар сонига ҳам таъсир қилади. **ГЛОТ-С** терминологиясининг маълумотлар базасида ижтимоий-сиёсий терминлар сони **500 (T = 500)** бўлганида, квота сўзи жадвали қуйидагича:

7- жадвал

α	υ	ρ	α	υ	ρ	α	υ	ρ
1	181	181	8	4	32	16	1	16
2	66	132	9	3	27	19	1	19
3	32	96	10	6	60	20	1	20
4	19	76	11	1	11	26	1	26
5	8	40	12	1	12	27	2	27
6	4	24	13	4	52	37	1	37
7	5	35	15	1	15	49	1	49

[1] Turney P.D. Coherent Keyphrase Extraction via Web Mining // Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico, 2003. – P. 434–439.

在这种情况下， $W = \sum \upsilon = 342$

并且 $P = \sum \rho = 987$

因此 $\Phi = P / W = 987/342 = 2.89$

当系统中的术语数为 1000 条（T = 1000）时，单词的频率表如下：

8- *жадвал*

α	υ	ρ	α	υ	ρ	α	υ	ρ
3	54	162	14	2	28	29	1	29
4	36	144	15	2	30	33	1	33
5	19	95	17	2	34	37	1	37
6	16	96	19	1	19	48	1	48
7	12	84	20	1	20	51	1	51
8	10	80	21	1	21	52	1	52
9	6	54	22	1	22	64	1	64
10	6	60	23	1	23			
11	2	22	24	1	24			

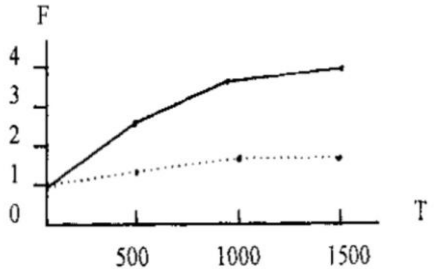
9- *жадвал*

α	υ	ρ	α	υ	ρ	α	υ	ρ
1	295	295	12	6	72	25	1	25
2	103	206	13	2	26	26	1	26

Таҳлил ва натижалар. Жадвалдан кўриниб турибдики, тизимдаги ижтимоий-сиёсий терминлар сонининг кўпайиши билан, расмда кўрсатилгандек, композитсион сўзларнинг частотаси ҳам шунга қараб ошади.[1]

（从表中可看出，随着系统中术语数的增加，单词的术语构成频率也相应地增加，图示如);

10- жадвал



Юқоридаги расмда, кесилган чизиқ Э тизимининг иқтисодий кўрсаткичининг ўзгаришини ва қаттиқ чизиқ ***Ф*** сўзининг терминлари таркибий қисмларининг частотасининг ўзгаришини билдиради.

[1] Sato S., Sasaki Y. Automatic Collection of Related Terms from the Web // The Companion Volume to the Proceedings of 41st Annual Meeting of the ACL, Sapporo, Japan, 2003. – P. 121–124.

Агар ***T*** терминлар сони бир хил бўлса, компонент сўзининг қиймати тизимнинг иқтисодий кўрсаткичидан кам бўлмайди. Ижтимоий-сиёсий терминлар сони **T = 1** бўлса ва тизимда биттагина сўз бўлса, **Ф га Э = тенг** келади. Бошқа ҳолларда, ***Ф*** ҳар доим Э дан каттароқдир.

Юқоридаги учта частота жадвалларидан ***а*** сўзининг мутлақ частотаси ошиши билан ***а*** сўзининг ҳар хил сўзларидаги ***в*** сони бир хил бўлганлиги аниқ бўлади. мутлақ частота мос равишда камаяди.

Ушбу муносабатни қуйидаги диаграмма билан тавсифлаш мумкин. [1]

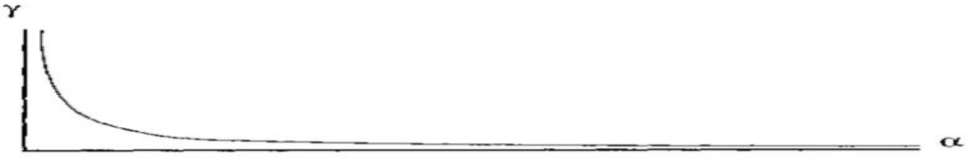
在上图中，虚线表示系统的经济指数E的变化情况，实线表示单词的术语构成频率 **Ф** 的变化情况，如果术语数 **T** 相同，单词的术语构成频率 **Ф** 的值不小于系统的经济指数E的值，即 **Ф>Э**。

仅当术语数 **T = 1**，系统中只有一个单词时，**Ф=E**，在其他场合，Ф永远大于E。

从上面三个频率表中还可看出，随着单词绝对频率 ***а*** 的增加，具有同一绝对频率的不同的单词的数目 ***в*** 相应地减小，

这种关系可用下图来描述:

11- жадвал



Бу шуни кўрсатадики, терминология тизимида юқори частотали сўзлар турли хил сўзларнинг умумий сонининг озгина қисмини эгаллайди, аммо улар кўп сонли терминлар бўлиши мумкин. *Масалан,* **T** терминлар сони **1510** бўлса, мутлақ частота **10** га тенг бўлган **62** та юқори частотали сўзлар мавжуд, аммо уларнинг пайдо бўлиши **1342** сўзни ташкил қилади.

Ушбу юқори частотали сўзлардан ташкил топган ишлайдиган сўзларнинг умумий сони сўзларнинг умумий сонидан атиги ***41,4%*** ни ташкил қилади. Терминология тизимида юқори частотали сўзлар қанчалик кўп бўлса, тизимдаги сўзлар терминларининг частотаси шунчалик юқори бўлади. [2]

这说明，在一个术语系统中，高频词只占了不同单词总数的一，
小部分，而它们却能构成大量的术语。
例如，当术语数T为 **1510** 条时，绝对频率人于**10**的高频词只有 **62** 个
而它们的出现次数却是**1342**词次，
由这些高频词构成的运行单词总数占了全部的运行单词总数的***41.4%***。
术语系统中的高频词越多：则该系统中单词的术语构成频率也就越高。

Хулоса. Сиёсий нутқни таржима қилиш муаммолари, аввало, сиёсий коммуникацияларни кўриб чиқишни талаб қилади, бу каби турли хил таркибий қисмлар мавжудлиги билан ажралиб туради: сиёсий вазият, муаллифнинг муносабати ва ҳк. уни аниқ ҳаракатлар қилиш. Ижтимоий-сиёсий терминларнинг частотасини ҳисоблаш учун, ва уни ҳисоблаш усулларни

[1] Baroni M., Bernardini S. BootCaT: Bootstrapping Corpora and Terms from the Web // Proceedings of LREC 2004. Lisbon: ELDA, 2004. – P. 1313–1316.
[2] Peñas A., Verdejo F., Gonzalo J. Corpus-Based Terminology Extraction Applied to Information Access // Proceedings of Corpus Linguistics 2001, Lancaster University, UK, 2001. – P. 458–465.

очиб бериш, амалга ошириш, муаллиф керакли натижага эришишни таъминлаш учун матн тузишда тегишли методлардан фойдаланади.

Юқорида таъкидлаб ўтилганидек, сиёсий матнларда турли хил лингвистик ва лисоний бўлмаган манбалардан фойдаланилади, масалан: метафора, метонимия, фразеологик бирликлар, турли хил услубий хусусиятлар сўз бирикмаларидан фойдаланиш, тарихий ва маданий дақиқаларни жалб қилиш ва бошқалар. Бундай композитсион тузилишга эга бўлган матнлар, албатта, таржима учун катта қийинчиликларни келтириб чиқаради. Бундай матнларнинг муваффақиятли таржимаси, аввало, таржимоннинг лингвистик тайёргарлигига боғлиқ.

Чет тили элементларини ассимилятсия қилиш жараёни улар таржимонлар томонидан тилга киритилган пайтдан бошлаб бошланганлиги сабабли, махсус чет эл номларини таржима қилиш техникаси ҳақида бир неча сўз айтиш ўринли. Тушунарсиз ассотсиатив маънога эга сўзлар ва ибораларни таржима қилишда, шунингдек, ҳақиқат номларини таржима қилишда, транскрипцияда, камроқ тез-тез транслятсия қилиш, излаш ва тушунтириш таржимаси (чет тилидаги сўз ёки иборанинг маъносини она тили ёрдамида мотиватсияни сақламасдан ўтказиш) ва шакл) ишлатилиши мумкин. Кузатиш, транскрипсия ва транслятсия пайтида баъзида изоҳларга мурожаат қилиш керак бўлади. Бирлаштирувчи маънонинг изоҳли таржимаси ва изини алмаштиришнинг бир тури бўлган техникани ҳам қўллаш мумкин.