Îá îäíîì ìåòîäå ñòàòèñòè÷åñêîé ôèëüòðàöèè òåêñòîâîé èíôîðìàöèè

Àâòîð Ñòàíèñëàâ Âàñèëüåâè÷ Ìî÷åíîâ
18.07.2008 ã.
Ïîñëåäíåå îáíîâëåíèå 20.07.2008 ã.

Òåêñò ïå÷àòíîãî èçäàíèÿ â ôîðìàòå PDF

Ââåäåíèå

Ïðîáëåìà àíàëèçà è ñèíòåçà òåêñòîâûõ äîêóìåíòîâ ïîñâÿùåíî çíà÷èòåëüíîå êîëè÷åñòâî ðàáîò [Ìàòåìàòè÷åñêèå 1977; Ôîìåíêî 1980; ïîëíîòåêñòîâûõ äîêóìåíòîâ, àâòîìàòè÷åñêîé êëàññèôèêàöèåé è èäåíòèôèêàöèåé òåì äîêóìåíòîâ, àâòîìàòè÷åñêèì ðåôåðèðîâàíèåì,

Ñòàòèñòè÷åñêèå ìåòîäû äîñòàòî÷íî õîðîøî çàðåêîìåíäîâàëè ñåáÿ ïðè ïîñòðîåíèè ïîèñêîâûõ ñèñòåì, âûäåëåíèè êëþ÷åâûõ ñëîâ è ñ

Â òî æå âðåìÿ ïðè ðåøåíèè çàäà÷ àíàëèçà è ñèíòåçà òåêñòîâîé èíôîðìàöèè, âîçíèêàþùèõ ïðè ïîñòðîåíèè èíôîðìàöèîííûõ ñèñòåì, â ñèíòàêñè÷åñêîãî è ñåìàíòè÷åñêîãî àíàëèçà.

Â äàííîé ñòàòüå ïðîâîäèòñÿ îáçîð ñòàòèñòè÷åñêèõ ìåòîäîâ àíàëèçà òåêñòîâ íà åñòåñòâåííûõ ÿçûêàõ, à òàêæå ïîêàçàíû âîçìîæíî

1. Êðàòêèé îáçîð ñòàòèñòè÷åñêèõ ìåòîäîâ èñïëåäîâàíèÿ òåêñòîâîé èíôîðìàöèè

Øèðîêîå ðàñïðîñòðàíåíèå ñòàòèñòè÷åñêèå ìåòîäû íàøëè â ãóìàíèòàðíûõ îáëàñòÿõ çíàíèé, â ÷àñòíîñòè, â èñòîðèè è ëèòåðàòóðîâ àâòîðñòâà ëèòåðàòóðíîãî ïðîèçâåäåíèÿ «Òèõèé Äîí» Ì. Â. Øîëîõîâà [Ôîìåíêî 1985]. Áîëüøèíñòâî ýòèõ ìåòîäîâ èñïîëüçóþò îáíè è òå

Ïðè ýòîì ïîä àâòîðñêèì èíâàðèàíòîì ïîíèìàþòñÿ íåêîòîðûå ìîæåñòâî ñòàòèñòè÷åñêèõ îïèñàíèé, ñ ïîìîùüþ êîòîðûõ îõóò ïðèíàäëåæíîñòü ðàçíûå çíà÷åíèÿ äëÿ ïðîèçâîëüíîé ðàçíîãî âðåìåíè.

Îáùåé ìîäåëüþ ôóíêöèîíèðîâàíèÿ ëèòåðàòóðíûõ òåêñòîâ, ñèëüíî çàòðóäíÿþùåé ôîðìàëèçàöèþ, ÿâëÿåòñÿ ðàçäåëåíèå íà ñòèëè, ÷àñòîòà óïîòðåáëåíèÿ ðàçëè÷íûõ ñëîâ ìîæåò ñëóæèòü íåêîòîðûì ïîêàçàòåëåì ñòèëÿ àâòîðà.

Î÷åâèäíî, ÷òî êîëè÷åñòâåííàÿ îöåíêà èíäèâèäóàëüíûõ îòëè÷èòåëüíûõ îñîáåííîñòåé ïðîèçâåäåíèÿ òîãî èëè èíîãî àâòîðà âåñüìà íåò

Ïðè âûáîðå òåõ èëè èíûõ ñòàòèñòè÷åñêèõ õàðàêòåðèñòèê, èñïîëüçóåìûõ â êà÷åñòâå èíäèêàòîðíûõ, íåîáõîäèìî, ÷òîáû îíè óäîâëåòâ

– èíòåãðàëüíîñòè (îáîáùåíèå ïî ñîâîêóïíîñòè ãðóïïû ïîêàçàòåëåé);

– ïîñòïýíòàó (äëÿ ïîðåäåëåííîãî àâòîðà íà ãðóïïå ïðîèçâåäåíèé);

– äèàïàçîíó èçìåíåíèé (äëÿ ãðóïï àâòîðîâ).

Òàêîå ñî÷åòàíèå âñåõ òðåõ ïåðå÷èñëåííûõ óñëîâèé ïîçâîëÿò îãîâîðèòü î íàëè÷èè íåêîòîðîãî àâòîðñêîãî èíâàðèàíòà.

Â êà÷åñòâå êîëè÷åñòâåííûõ õàðàêòåðèñòèê òåñòîâ ïðåäëàãàþòñÿ ñëåäóþùèå:

– äëèíà ïðåäëîæåíèé (ñðåäíå÷èñëî ñëîâ â ïðåäëîæåíèè, ïîäñ÷èòàííîå äëÿ êàæäîé âûáîðêè);

– äëèíà ñëîâ (ñðåäíå÷èñëî ñëîâ â ñëîâå, ïîäñ÷èòàííîå äëÿ êàæäîé âûáîðêè);

– ÷àñòîòà óïîòðåáëåíèÿ ñëóæåáíûõ ñëîâ ïðåäëîãîâ, ñîþçîâ, ÷àñòèö;

– ÷àñòîòà óïîòðåáëåíèÿ ñóùåñòâèòåëüíûõ;

– ÷àñòîòà óïîòðåáëåíèÿ ãëàãîëîâ;

– ÷àñòîòà óïîòðåáëåíèÿ ïðèëàãàòåëüíûõ;

– ÷àñòîòà óïîòðåáëåíèÿ ïðåäëîãà «â»;

– ÷àñòîòà óïîòðåáëåíèÿ ÷àñòèöû «íå»;

– êîëè÷åñòâî ñëóæåáíûõ ñëîâ â ïðåäëîæåíèè (ñðåäíåå ÷èñëî ñîþçîâ, ïðåäëîãîâ è ÷àñòèö â ïðåäëîæåíèè äëÿ êàæäîé âû

Ñëåäóåò îòìåòèòü, ÷òî â óêàçàííûõ ðàáîòàõ [Õàðèí] ïðèâåäåíû ñòàòèñòè÷åñêè õàðàêòåðèñòèêè èñïîëüçîâàëèñü ëèøü äëÿ
ãðóïï ââòîðîâ. Â òî æå âðåìÿ ïðè ðåøåíèè çàäà÷ àíàëèçà âàæíûì ÿâëÿåòñÿ ñîêðàùåíèå îáúåìà èñõîäíîé òåêñòîâîé èíôîðìàöèè áåç

Íèæå ïðåäëàãàåòñÿ àëãîðèòì ôèëüòðàöèè, îñíîâàííûé íà èñïîëüçîâàíèè îäíîé èç âûøå ïîèñàííûõ õàðàêòåðèñòèê.

2. Àëãîðèòì ôèëüòðàöèè

Â ðàáîòàõ [Ìî÷åíîâ è äð. 2004] ïðè ïîñòðîåíèè ñòàòèñòè÷åñêîé èíôîðìàöèîííîé ìîäåëè èñïîëüçîâàëèñü òåêſòû ïî ýêſíîìè÷åſêé, ýêſê
10 ñòðàíèö ìàøèíîïèñíîãî òåêſòà ôîðìàòà À4, ñîäåðæàùèõ ïîðÿäêà 300 ïðåäëîæåíèé êàæäûé).

Ïðåäìåòîì äàííûõ èññëåäîâàíèé ÿâëÿåòſ âûÿâëåíèå îáîáùåííûõ ñòàòèſòè÷åſêèõ õàðàêòåðiſòèſê ðàſïðåäåëåíèÿ êîëè÷åſòâà ſë
ðåøåíèå ſëåäóþùèõ çàäà÷:

– óìåíüøåíèå îáùåãî îáúåìà àíàëèçèðóåìîãî òåêñòà ïðè ñîõðàíåíèè åãî ñåìàíòè÷åñêîé ñîñòàâëÿþùåé;

– àâòîìàòè÷åñêîå ðåôåðèðîâàíèå;

– âûÿâëåíèå ñïåöèôè÷åñêèõ ôðàãìåíòîâ òåêñòà, óäîâëåòâîðÿþùèõ çàäàííûì êðèòåðèÿì;

– ïîäãîòîâêà òåêñòà äëÿ äàëüíåéøèõ ýòàïîâ àíàëèçà.

Âûÿâëåííûå ñòàòèñòè÷åñêè õàðàêòåðèñòèêè ìîãóò áûòü èñïîëüçîâàíû è ïðè ðåøåíèè çàäà÷ ñèíòåçà òåêñòà.

Â ïðîöåññå èññëåäîâàíèé áûëè âûïîëíåíû ýêñïåðèìåíòû, íàïðàâëåííûå íà èçó÷åíèå ðîëè ïðåäëîæåíèé ðàçíîé äëèíû. Íà ðèñ. 1 ïîêàç

Ðèñ. 1. Ðàñïðåäåëåíèå ïðåäëîæåíèé ïî êîëè÷åñòâó ñëîâ â ïðåäëîæåíèè

Íà ãðàôèêå ïî îñè îðäèíàò ïîëîæåíî êîëè÷åñòâî ïðåäëîæåíèé, à ïî îñè àáñöèññ — èíòåðâàëû ãðóïï ïî ÷èñëó ñëîâ â ïðåäëîæ

Èññëåäîâàíèÿ ïîêàçàëè, ÷òî äëÿ äàííûõ òåêñòîâ áîëüøàÿ ÷àñòü ïðåäëîæåíèé ïðèõîäèòñÿ íà ïðåäëîæåíèÿ ñðåäíåé äëèíû: îò 10 äî
ñìûñëîâîé èòîã àáçàöà. Ïðåäëîæåíèÿ áîëüøîé äëèíû (îò 17 ñëîâ è áîëåå) ëèáî óòî÷íÿþò íåêîòîðûé ñìûñë ïðåäûäóùåãî ïðåäëîæåíè

Íàïðàâëåííàÿ ôèëüòðàöèÿ ïðåäïîëàãàåò èñïîëüçîâàíèå ñòàòèñòè÷åñêèõ õàðàêòåðèñòèê ðàñïðåäåëåíèé, òàêèõ, êàê ìàòåìàòè÷

Äëÿ öåëåé íàïðàâëåííîé ôèëüòðàöèè ìîãóò áûòü èñïîëüçîâàíû è äðóãèå îñîáùåííûå õàðàêòåðèñòèêè, î êîòîðûõ øëà ðå÷ü âûøå.

Çàêëþ÷åíèå

Ñòàòèñòè÷åñêèå ìåòîäû àíàëèçà ðóññêîÿçû÷íûõ òåêñòîâ ìîãóò ñ óñïåõîì ïðèìåíÿòüñÿ äëÿ ðåøåíèÿ ðàçíîîáðàçíûõ çàäà÷ îáðàáîòê
ôèëüòðàöèÿ ïîçâîëÿåò îïåðèðîâàòü ýëåìåíòàìè òåêñòà, íàïðèìåð, ãëàâîé, ñòðàíèöåé, àáçàöåì.

Summary

The paper presents a review of the statistic methods of analysis of texts in natural languages and shows the possibilities of filtration of the text information on the basis of one of the most informative statistic text characteristics. There is developed a method of filtration of the texts in the Russian language that can be applied to the solution of various tasks of text processing like the decrease of the volume of textual information, writing essays, determination of the semantic component of the text units etc.

Ñïèñîê ëèòåðàòóðû

Ìàòåìàòè÷åñêèå 1977   Ìàòåìàòè÷åñêèå ìåòîäû â èñòîðèêî-ýêîíîìè÷åñêèõ è èñòîðèêî-êóëüòóðíûõ èññëåäîâàíèÿõ.   Ì., 1977.

Ìàòåìàòè÷åñêèå 1985   Ìàòåìàòè÷åñêèå ìåòîäû è ÝÂÌ â èñòîðè÷åñêèõ èññëåäîâàíèÿõ.   Ì., 1985.

Ìî÷åíîâ è äð. 2005   Ìî÷åíîâ, Ñ. Â. Ïðèìåíåíèå ñòàòèñòè÷åñêèõ ìåòîäîâ ïðè àíàëèçå òåêñòîâîé èíôîðìàöèè / Ñ. Â. Ìî÷åíîâ, À. Ì. Áëåä

Íîñîâñêèé è äð. 1989   Íîñîâñêèé, Ã. Â. Ñòàòèñòè÷åñêèå äóáëèêàòû â óïîðÿäî÷åííûõ ñïèñêàõ ñ ðàçáèåíèåì / Ã. Â. Íîñîâñêèé, À. Ô
èññëåäîâàíèÿ.   Ì., 1989.   Ñ. 138–148.   (Íàó÷. ñîâåò ïî êîìïëåêñíîé ïðîáëåìå «Êèáåðíåòèêà», ÀÍ ÑÑÑÐ).

Ôîìåíêî 1980   Ôîìåíêî, À. Ò. Íåêîòîðûå ñòàòèñòè÷åñêèå çàêîíîìåðíîñòè ðàñïðåäåëåíèÿ ïëîòíîñòè èíôîðìàöèè â òåêñòàõ ñî øêàëîé

Ôîìåíêî 1983 Ôîìåíêî, À. Ò. Àâòîðñêèé èíâàðèàíò ðóññêèõ ëèòåðàòóðíûõ òåêſòîâ // Ìåòîäû êîëè÷åſòâåííîãî àíàëèçà òåêſòîâ íà
Ñ. 86–109.

Ôîìåíêî 1985 Ôîìåíêî, À. Ò. Èſòîðìàòèâíûå ôóſêöèè è ſâÿçàſíûå ſ íèì ſòàòſòè÷åſêèå çàêſſîìåðſſſòè / À. Ò. Ôîìåíêî // Sòàòſòè

Õàðèſ Õàðèſ, Í. Ï. Èſſſëäſâàſíèå ïðſſſöſïîâ ſåì àſſòſ÷åſêſãſ ïſſſêà òåſſòſâſſé èſſſðſàöſè íà îſſſåå èſſïſëüçſâàſſÿ èſòåëåêòóàëüſſ