Комитет науки Министерства образования и науки
Республики Казахстан
РГП «Институт информационных и вычислительных технологий»
КН МОН РК

25 лет
Независимости
Республики Казахстан

25 лет
Институту
информационных и
вычислительных
технологий

МАТЕРИАЛЫ

Международной научной конференции
«Информатика и прикладная математика»
(«Computer science and Applied Mathematics»),
посвященной 25-летию Независимости Республики Казахстан и
25-летию Института информационных и
вычислительных технологий

Часть II

г. Алматы, 21-24 сентября 2016 года

Алматы
2016

Содержание

ГИПЕРГРАФ НЕГІЗІНДЕ КІЛТТІК СӨЗДЕРДІ ШЫҒАРУ

Солтангельдинова М.К., Мансурова М.Е., Бердибеков С.

Әл-Фараби атындағы Қазақ Ұлттық Университеті, Қазақстан

Аннотация. *Ұсынылып отырған жұмыста кілттік сөздерді шығару алгоритмі, соның ішінде семантикалық байланысқан сөздерді шығару технологиясы қарастырылған. Алгоритм графтың ерекше түрі – гиперграфқа негізделеді. Сонымен қатар, «Кездейсоқ кезу» әдісін қолданыла отырып, өңделетін құжаттың салмағы қарастырылады. Бұл семантикалық байланысқан сөздерді шығаруда, яғни өңделетін ғылыми құжаттардың барлық ерекшеліктерін ескеруді қамтамасыз етеді. Нәтижесінде алынатын семантикалық байланысқан сөздер, фактографилық ақпараттық жүйесін құруда жалғасын табады.*

Кіріспе

Қазіргі таңда ақпарат көлемі өте үлкен жылдамдықпен ұлғаюда. Үлкен көлемді мәтіндерді оқу және сондай үлкен массивті деректерден ақпарат іздеу тиімсіз болып келеді. Сол себепті деректерді өңдеу жұмыстары алға шығып отыр. Мұндай деректерді өңдеу ақпараттық-аналитикалық іздеу жүйелерінің негізі болып табылады. Деректерді өңдеу, талдау бүгінгі таңда әр түрлі салаларда жиі қолданылады. Мәтіндерден қажетті білім шығара білу - талдау, өңдеу жұмыстарының жаңа бағыты болып есептеледі. Бұл деректерді өңдеу барысында семантикалық байланысқан сөздерді шығару есебін өзекті етіп отыр. Шығарылатын сөздер/ сөз тіркестері кейін көп мақсатта қолданылуы мүмкін. Сондықтан да кілттік сөздер, сөз тіркестерін шығарудың тиімді алгоритмін құру және зерттеу маңызды есептердің бірі.

Ұсынылып отырған алгоритмнің негізгі методикалық сипаттамалары жұмыстың теориялық мәселелер бөлімінде, ал практикалық жүзеге асырылуы тсептеуіш тәжірибе бөлімінде сипатталынған.

Жұмыстың мақсаты – фактографиялық ақпараттық жүйесін құру мақсатында семантикалық байланысқан сөздерді шығару. Осы мақсатқа жету негізінде келесі есептер алға қойылды:

1. Мәтіндерді морфолгиялық талдаудан өткізу.
2. Мәтіндерді гиперграф түрінде модельдеу, «Кездейсоқ кезу» процессі негізінде Марков шынжыры әдісін қолдану.
3. Граф төбелерінің рангтерін есептеу, саптау.
4. Семантикалық сөздерді шығару.

Жұмыстың тәжірибелік маңызы – деректерді өңдеп, нәтижесінде кілттік сөздердің алынуы кейін фактографиялық іздеу жүйелерін ұйымдастыруға арналған, құрылатын онтолияның пәндік аймағында сипатталатын, болмыстың мүмкін болатын мәндері ретінде қолданылуы.

1 Мәтіннің морфологиялық белгілеу

Мәтіндерді морфологиялық белгілеу табиғи тілде мәтіндерді өңдеудің ең маңызды қадамы. Мәтіндердің морфологиялық белгілеуі сөз формаларының белгілеуінен және әрбір сөз формасына лексемалық сипаттамалар мен оның грамматикалық белгілер жиынтығын меншіктеуден тұрады.

Мәтіндегі әрбір сөзге құрылатын морфологиялық ақпарат төрт «жолдан» немесе белгілеулер тобынан тұрады:

1. Сөз формасы жататын лексема (лексеманың «сөздік жазбасын» және қай сөз табы екендігін көрсетеді).
2. Лексеманың грамматикалық белгілер жиынын немесе сөз классификациялаушы характеристикалары.
3. Сөз формасының грамматикалық белгілер жиыны немесе сөз түрлендіруші сипаттамалары (мысалы, зат есім септігі, етістік түрі яғни жекеше, көпше түрлері).
4. Грамматикалық форманың стандартты еместігі, орфографиялық бұрмалануы туралы ақпарат.

Мәтіндерді морфологиялық белгілеу бірнеше кезеңнен өтеді:

1. Парсинг. Мәтіндерді өңдеудің бұл кезеңінде, сөздерге талдау жүргізіледі. Парсер әрбір сөзге анализ жасап, талдау нәтижесін шығарады.
2. Фильтрация. Талдауды тазалаудан өткізеді және талдау нұсқаларын жояды.
3. Қолдан омонимияны шешу процессі.

2 Гиперграф негізінде кілттік сөздерді шығару

Гиперграф – әрбір қабырғамен екі немесе одан да көп ішкі төбелер жиыны байланыса алатын жалпылама граф.

Айталық, бізде n документтен тұратын мәтін бар деп қарастырайық.

$$D = (d_1, d_2, ...d_n) \quad (1)$$

Мәтін анализі кезінде, документ ретінде, негізінде, жеке алынған сөйлемдер қарастырылады. Келесі белгілеулерді енгізейік:

$$W = (w_1, w_2, ...w_n) \quad (2)$$

- біздің мәтініміздің сөздігі.

Мәтінідер алдын ала морфологиялық талдаудан өткендіктен, әрбір сөздің морфологиялық сипаттамалары белгілі [2].

2.1 Мәтінді гиперграф түрінде модельдеу.

Талдау жүргізілетін мәтінді гиперграф түрінде модельдеу өте ыңғайлы әрі көп қолданылатын әдістің бірі. Талдау жасалатын мәтінімізді, HG(V, E) гиперграф түрінде, V төбелер жиыны және E гиперқабырғалар жиыны ретінде көрсетуге болады. E гиперқабырға V төбелер ішкі жиыны болып келеді, мұндағы $\bigcup_{e \in E} e = V$. Гиперграфтағы $v \in V$ төбелер ол мәтін сөздері, ал $e \in E$ гиберқабырғалар біздің мәтініміздің документтері.

$HG = R^{|V| \times |E|}$ арқылы біздің гиперграфтың араласын матрицасын белгілейік:

$$h(v,e) = \begin{cases} 1, & егер \quad v \in e \\ 0, & егер \quad v \notin e \end{cases} \quad (3)$$

$HG(V, E, w)$ - өлшенген гиперграф болсын. Мұндағы $w : E - R^+$ - гиперқабырға салмағы.

Біздің жағдайда, гиперграфтың төбе және гиперқабырға дәрежесі келесі формуламен анықталады:

$$d(v) = \sum_{e \in E} w(e) h(v,e) \quad (4)$$

$$\delta(e) = \sum_{v \in V} h(v,e) = |e| \quad (5)$$

D_e және D_v гиперқабырғалар және төбелердің дәрежесін сипаттайтын диагональді матрица.

2.2 Гиперграф төбелерінің салмағын модельдеу

Мәтіндерді өңдеуде әрбір сөз маңыздылығын өлшеу, яғни сөздің мәтін TF-IDF өлшемі мәтіндер анализі және ақпараттық іздеу есептерінде жиі қолданылады. Мұндай өлшем көбінесе коллекция документтерін сандық вектор түрінде көрсетуде қолданыс табады, яғни қандай да бір сөздер жиынынан әрбір сөздің қолданылу маңыздылығын көрсетеді.

Біздің қарастырып отырған есебімізде, граф төбелерінің салмағы, яғни өңделетін мәтіндердің сөздерінің салмағын TF-IDF әдісіне негіздеп, келесі формуламен анықтаймыз:

$$w(v_i)_{tf-idf} = \frac{tf(v_i)}{N_w} * \log \frac{N}{df(v_i)} \quad (6)$$

Мұндағы:

- $tf(v_i)$ - документте жиі кездесетін термин,
- N_w - документке кіретін барлық сөздер суммасы
- N - D документтер коллекциясындағы документтер саны
- $df(v_i)$ - v_i термині кездесетін D-дағы документтер жиыны.

Гиперқабырға салмағы ретінде ,мысалы, документ танымалдылығын алуға болады. Осындай ерекшеліктер ескеріліп, келесі формула арқылы гиперқабырға салмағы есептеледі:

$$w(d_i) = \lambda R_{social}(d_i) + (1 - \lambda) R_{time}(d_i) \quad (7)$$

Мұндағы:

- λ - $0 < \lambda < 1$ аралығындағы тегістеу параметрі (сглаживание)
- R_{social} - әлеуметтік саптау (социальное ранжирование)
- R_{time} - уақыттық әсер ету өлшемі

Бұл параметрлер келесі формулалармен анықталады:

$$R_{time}(d_i) = Q^{(c-y_i)/24} \quad (8)$$

Бұл жердегі c және y_i ағымдағы уақыт және d_i құжатының жарияланған күні. Оны 24-ке бөлу арқылы жарияланған уақыт пен ағымдағы уақыт арасындағы айырмашылықты күндер санынан көреміз. Q 0 мен 1 аралығындағы өшу параметрі болып табылады. Q -ды 0,5 деп алу қабылданған.

$$R_{social}(d_i) = \frac{s_i + 1}{\sum_e s_e + 1} \quad (9)$$

Мұндағы, s_i d_i документі үшін социалды ерекшеліктер саны. $\sum_e s_e$ барлық документтедегі социалды ерекшеліктер қосындысы.

Осы ерекшеліктердің бәрі қамтылып, есептеу барысында гиперқабырға салмағының формуласын анықтайды. Бұл гиперқабырға және төбелер салмағы кейін есептеулерде өз қолданысын табады.

2.3 «Кездейсоқ кезу» процесі. Марков тізбегі

Кездейсоқ кезу – дискретті уақыт мезеттерінде қадамдардың кездейоқ өзгеру процессінің математикалық моделі. Сонымен қоса, әрбір қадамдағы өзгерулер алдындағылардан және уақыттан тәуелсіз деп ұйғарылады. Бұл модель әртүрлі сфераларда қолданылады, сондай – ақ, мұндай модель шынай процесстің айтарлықтай жеңілдеуі болып табылады.

Бірөлшемді дискретті кездейсоқ кезу – бұл дискретті уақытты $\{Y_n\}_{n\geq 0}$ кездейсоқ процесс және оның түрі келесідей:

$$Y_n = Y_0 + \sum_{i=1}^{n} X_i \qquad (10)$$

Мұндағы:

- Y_0 – бастапқы күй.

- $X_i = \begin{cases} 1, & p_i \\ -1, & q_i = 1 - p_i \end{cases}$, $0 < p_i < 1$, i$\in$ N

- Y_0, X_i, i =1,2… кездейсоқ шамалар бірге тәуелсіз.

Бірөлшемді дискретті кездейсоқ кезу бүтін күйлері бар *Марков тізбегі* болып табылады. Оның бастапқы үлестірілімдері X_0 кездейсоқ шаманың ықтималдық функциясы арқылы беріледі, ал көшудің ықтималдықтар матрицасының түрі келесідей:

$$P \equiv (p_{ij})_{i,j\in Z} = \begin{pmatrix} \ddots & \ddots & \ddots & & \\ & q_{-1} & 0 & p_{-1} & \\ & & q_0 & 0 & p_0 \\ & & & q_1 & 0 & p_1 \\ & & & & \ddots & \ddots & \ddots \end{pmatrix}$$

Яғни:

$p_{i,i+1} \equiv \mathrm{P}(X_{n+1} = i+1 \,|\, X_n = i) = p_i$,

$p_{i,i-1} \equiv \mathrm{P}(X_{n+1} = i-1 \,|\, X_n = i) = q_i$, $i \in \mathrm{Z}$,

$p_{ij} \equiv \mathrm{P}(X_{n+1} = j \,|\, X_n = i) = 0$, $|i-j| \neq 1$.

$\{X_n\}_{n\geq 0}$ дискретті кездейсоқ шамалар тізбегі қарапайым Марков тізбегі деп аталады, егер: $\mathrm{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathrm{P}(X_{n+1} = i_{n+1} \mid X_n = i_n)$ Осылайша, қарапайым жағдайда, Марков тізбегінің келесі күйлерінің шартты үлестірілімдері ағымдағы күйге ғана тәуелді және алдыңғы күйлердің барлығынан тәуелсіз.

$\{X_n\}$ кездейсоқ шаманың мәндер облысы –тізбек күйлерінің кеңістігі деп аталады, ал n - қадам нөмірі.

$P(n)$ матрицасы n -ші қадамдағы көшулер ықтималдығының матрицасы деп аталады:

$$P_{ij}(n) = \mathrm{P}(X_{n+1} = j \mid X_n = i) \qquad (11)$$

Ал вектор $p = (p_1, p_2, \ldots p_n)^T$, мұндағы $p_i = \mathrm{P}(X_0 = i)$ – Марков тізбегінің бастапқы үлестірілімі. Көшулер ықтималдығының матрицасы стохастикалық болатындығы айқын, яғни:

$$\sum_j P_{ij}(n) = 1, \forall n \in N \quad (12)$$

Марков тізбегі бірыңғай деп аталады, егер көшулер ықтималдығының матрицасы қадам нөмірінен тәуелсіз болса:

$$P_{ij}(n) = P_{ij}, \forall n \in N \quad (13)$$

Гиперграф төбелерін саптау үшін кездейсоқ кезу процессін гиперграфқа жалпылаймыз. Кездейсоқ кезу процессі графтағы төбелер арасындағы ауысулар, яғни әрбір t қадамды дискретті уақыттан кейін берілген төбеден көрші төбеге ауысу болып табылады. Біз төбелерді $\{s_1, s_2, ...s_n\}$күйлер жиыны ретінде қарастыра аламыз, ал ауысулар осы күйлердің ақырлы М Марков тізбегі болады. Ауысу ықтималдығы $P(u,v) = \Pr ob(s_{t+1} = v \mid s_t = u)$ түрінде есептеледі, бұл М Марков тізбегінің v–да t+1 уақытта болатындығын, t уақытта u-да болғандығын білдіреді. Біздің жағдайда, Марков тізбегі бірыңғай болып табылады, ауысу ықтималдығы t уақыттан тәуелсіз. Әрбір u төбесі үшін $\sum_v P(u,v) = 1$.

М тек бір ғана ауысу үшін есептелген ықтималдықтары бар бірыңғай болып табылады. Барлық жүріс үшін ауысу матрицасын $P \in \mathbb{R}^{|V|\times|V|}$ есептеуге болады. P ауысулар матрицасы сәйкес сондай ықтималдықтармен төбелер арасындағы серфер іс әрекетінің кездейсоқ түрде орын ауыстыруын көрсететін төбелер арасында ауысуларды толығымен қамтиды.

Қарапайым графтарда, кездейсқ кезу процессі түсінікті болып келеді, тек қана ықтималдығы анықталған мақсатты төбеге баратын қабырға таңдалады. Сонда да, гиперграф үшін бұл жағдай біраз өзгеше орын алады. Бұл гиперграф құрылымының өзгешелігінде. Мысалы, гиперграфта, гиперқабырғаның екіден көп $\delta(e) \geq 2$ төбелер нүктелері болуы мүмкін.

Гиперграфта кездейсоқ кезу процессін жалпылау үшін, біз гиперқабырғада бір бірімен инцидентті болатын екі төбе арасындағы ауысуды кезу түрінде модельдейміз. Асылында, кездейсоқ кезу процессі бір қадамның орнында екі қадамдық процесс: бірінші, кездейсоқ серфер ағымдағы u төбесімен инцидентті болатын e гиперқабырғасын таңдайды. Екінші, серфер $u, v \in e$ шартын қанағаттандыратын, таңдалған гиперграфтан мақсатты v төбесін таңдайды.

Кездейсоқ кезу гиперграфта жалпылама аталады, ал қарапайым графта кездейсоқ кезудің ерекше жағдайы бар. Ол яғни, қабырғаның бір ғана төбесінің болуы, ал гиперграфта көп төбелер арасынан таңдай аламыз. Гиперграфпен кездейсоқ кезу процессін Марков шынжыры арқылы анықтайтын болсақ, мұнда төбелер жиыны күйлер жиынын құрайтын болады. Әрбір t уақытты қадам сайын, серфер инцидентті гиперқабырғада басқа төбелерге орын ауыстырады.

Бұл жұмыста, біз кездейсоқ кезудің өлшенген гиперграфтағы жалпы анықтамасын анықтауға тырысамыз, яғни гиперқабырғамен қоса төбелердің де салмағы ескерілетін болады. Мұндай жағдайда, кездейсоқ кезу процессі гиперқабырға және төбелер салмағын қолдану арқылы кеңейе түседі. Төбе салмағын барлық инцидентті гиперқабырғалар арқылы анықтаймыз, ол ерекшеліктер векторы:

$$\vec{v_w} = \{w(v_{e1}), w(v_{e2}), ..., w(v_{d(v)})\} \quad (14)$$

Яғни, бізде v төбесі кездесетін әрбір e гиперқабырғасы үшін төбелердің әртүрлі салмағы бар. Кездейсоқ кезудің болжамалы процессін келесі түрде сипаттаймыз: u төбесінен бастап серфер гиберқабырға салмағына $w(e)$ пропорционалды болатын, u төбесімен инцидентті болатын e гиперқабырғасын таңдайды. Одан кейін серфер, дәл солай, гиперқабырғада төбе салмағына пропорционалды v төбесі таңдалады, яғни ағымдағы гиперқабырғада қарастырылып отырған салмақ.

Өлшенген гиперграфтың инцидентті матрицасын $H_w \in R^{|V|\times|E|}$ келесідей анықтайық :

$$h_w(v,e) = \begin{cases} w(v_e), & if \quad v \in e \\ 0, & if \quad v \notin e \end{cases} \tag{15}$$

Осылайша, гиперқабырға дәрежесін қайтадан анықтасақ:

$$\delta(e_w) = \sum_{v \in V} h_w(v,e) \tag{16}$$

Жоғарыда көрсетілген формулаларды қолдана отырып, P көшу (ауысу) матрицасын есептеуге болады:

$$P(u,v) = \sum\nolimits_{e \in E} w(e) \frac{h(u,e)}{\sum\limits_{\hat{e} \in E(u)} w(\hat{e})} \frac{h_w(v,e)}{\sum\limits_{\hat{v} \in e} h_w(\hat{v},e)} \tag{17}$$

Немесе матрицалық белгілеу:

$$P = D_v^{-1} H W_e D_{ve}^{-1} H_w^T \tag{18}$$

Мұндағы:

- D_v - төбенің өлшенген дәрежесінің диагональді матрицасы;
- H – гиперграф төбелерінің инцидентті матрицасы;
- W_e – гиперқабырға салмақтарының диагональді матрицасы;
- D_{ve} – гиперқабырғаның өлшенген дәрежесінің диагональді матрицасы;
- H_w – өлшенген графтыі инцидентті матрицасы.

Мұнда P көшу матрицасы стохастикалық болып табылады және әрбір жолдар қосындысы 1-ге тең [1].

P көшу матрицасын есептегеннен кейін, кездейсоқ кезудің π стационарлы үслетірілімін түсіне қажеттігі туындайды. Стационарлы үлестірілім $\vec{v}_0 \in R^{|V|\times 1}$ векторынан бастап есептелуі мүмкін. Ол 1/|V| қосындылары 1-ге тең ықтималдықар. Ең алдымен, P^T көшу матрицасын $\vec{v}_0$ баған векторына көбейту арқылы $\vec{v}_0 = P^T * \vec{v}$ алынады. Осылайша, $\vec{v}$ векторы өзгеруін тоқтатқанға дейін итерацияны жалғастырамыз. $\vec{v}$ ықтималдықты үлестіру векторын көше матрицасына көбейту, келесі үлестірім қадамын $\vec{x} = P^T * \vec{v}$ береді. x_i ағымдағы i төбесінде болу ықтималдылығы болсын. Онда $x_i = \sum_j p_{ij} v_j$, v_j серфердің j түйінінде алдын ал болуының ықтималдылығы, және p_{ij} j-дан i-ға көшу ықтималдылығы.

$\vec{v}$ векторының үлестірілім ықтималдылығы өзгеруі n қадамнан кейін тоқтайды, егер кездейсоқ кезу эргодикалық болса. Кездейсоқ кезу эргодикалық болады, егер келесі шарттар орындалса:

1 Кез келген екі $s_i, s_j \in M$ күйлері үшін тізбек аударылмайтын (неприводимый) болса, және олар келесі шартты қанағаттандыруы тиіс: $P(s_i, s_j) > 0$.

2 Тізбек апериодикалық болса, яғни әрбір күйдің ЕҮОБ $\{t : P_t(s_i, s_j) > 0\}$ ол 1.

Аударылмаушылық және апериодикалық қасиеттерді тұрақтандыру үшін PageRank алгоритмін қолдануға болады [1].

3 PageRank алгоритмі

PageRank алгоритмі - бұл сілтемелік саптау алгоритмдерінің бірі. Бұл алгоритм көбінесе гиперсілтемелермен байланысқан документтер коллекцияларына қолданылады.

Бұл алгоритмнің кездейсоқ кезу процесіне қолданудағы негізгі идея, бұл алгоритм телепортация идеясын қолданады, яғни кездейсоқ кезу процессін қайтадан жібереді. Ал бұл жағдай кезедейсоқ кезудің эргодикалық қасиеттері үшін тиімді. Телепортация өте аз ықтималдықты өшу факторы (damping factor) арқылы көрсетіледі. Бұл сондай –ақ графты аударылмайтындай жасауға кепіл береді, себебі кездейсоқ кезудің кез келген басқа бір түйіндерге телепортация жасау ықтималдылығы бар.

$$\vec{v}_{i+1} = \alpha P^T \vec{v}_i + (1-\alpha) * 1/n \qquad (19)$$

Мұндағы:

- α – өшу факторы, оны 0,85 деп аламыз.
- n - графтағы түйіндер саны.
- $\alpha P^T v_i$ кездейсоқ кезу бір инцидентті гиперқабырғадан таңдалынатынын білдіреді.

4 Шаблондар әдісі

Жалпы тұрғыда кілттік сөздерді шығару әдістері өте көп. Ондай әдістердің қатарына N-gram, TF-IDF, TextRank, POS chunking және т.б. Соның ішінде POS chunking әдісі салыстырмалы түрде жақсы нәтиже көрсетеді.

POS chunking әдісі ,негізінде, морфологиялық шаблондарға негізделген. Бұл кілттік сөздерді шығару кезінде, мәтін сөздерінің қатаң сөз таптары мен оның характеристикаларын бақылауды талап етеді. POS моделін аңдатылған деректерден шығаруға болады. Кілттік сөздерді шығару процессі кезінде, ең алдымен, мәтіндерге талдау жүргізіліп, әрбір сөздің морфологиялық сипаттамалары алынады. Сол сипаттамаларды қолдану арқылы, сәйкес шаблондар құрылып, кілттік сөздерді шығару жүзеге асырылады.

POS chunking әдісінің артықшылығы, деректерден мүмкіндігінше мағынасы бар кілттік сөздерді шығаруға тырысады, яғни, ол мағынасыз кілттік сөздер санын қысқартады

5 Есептеуіш тәжірибе

Жоғарыда нұсқалған алгоритмге тәжірибе жүргізу барысында Леонид Орловтың «Как создать электронный магазин в интернет» оқу құралы қолданылды. Бұл кітап толықтай 384 бетті қамтиды. Тәжірибе барысында ең алдымен, кітап толықтай морфологиялық талдаудан өткізілді. Бағдарлама Java тілінде жазылғандықтан,

морфологиялық талдау жасауда Java тіліне негізделген Apache Lucene кітапханасы пайдаланылды. Талдаудан өткізілгеннен кейін құжатты қамтитын барлық сөздер сипаттамасы жеке түрде сақталынып отырды. Бұл сипаттамалар семантикалық сөздер / сөз тіркестерін шығаруда, шаблондар әдісін қолдану кезінде қажет болып табылады.

Кіріс ақпарат өңдеуден өткізілген соң, ендігі сол деректермен жұмыс жасалынды. Өңделген деректерді гиперграфқа негізделп, қажетті есептеулерді жүргіздік, яғни, ол құжаттың өзінің салмағын, құжат мәтініне кіретін әрбір сөздің салмағын анықтау, байланысу ықтималдықтарын есептеу. Бұл жағдайда, гиперграфтың өлшенген және өлшенбеген түрлері де қарастырылды.

Алгоритмнің ең маңызды бөлігі, сөздер арасындағы байланысты анықтау, сөздер рангін есептеу, «Кездейсоқ кезу» моделіне негізделді. Кездейсоқ кезу моделі қарапайым Марков тізбегінде көрсетілді. Яғни, тізбектің күйлері ретінде граф төбелері, қарастырып отырған сөздер жиыны алынды. Семантикалық байланысты анықтау кездейсоқ түрде, бір күйден екінші күйге көшу арқылы көшу ықтималдықтарын есептеу көмегімен анықталды. Көшу ықтималдығы, құжат мәтініне кіретін әрбір сөздің барлық сөздермен байланысу ықтималдығын көрсетеді. Ықтималдық жоғары болған сайын, байланысу мүмкіндігі де жоғары.

Кездейсоқ кезу процессінен кейін стационарлы үлестірім мәнін есептеу жүргізілді. Бұл мән PageRank алгоритміне негізделді, және де бұл есептеу бізге сөздің рангін береді. Сол ранг аркылы кілттік сөздерді анықтауға болады. Ранг жоғары болған сөздер, құжат аясында жиі кездесетін кілттік сөздер болу мүмкінідігі бар. Ранг есептеу кезінде (19) формулада көрсетілген, α өшу факторы мәні ретінде 0,89 мәнін алу келісілді. Осылайша, әрбір сөздің рангтері анықталды.

Кілттік сөздер / сөз тіркестерін шығару мақсатына қол жеткізуде шаблондар әдісі қолданылды. Шаблондар әдісі негізінде өңделетін тілдің құрылымына байланысты болып табылады, және де сол бойынша шаблондар анықталады. Біздің жағдайда, орыс тілді құжаттар қарастырылғандықтан, сәйкес шаблондарды анықтап алдық. Ол шаблондар келесідей болды:

[прилагательное + существительное];
[причастие + существительное];
[существительное + существительное, Род.падеж];
[существительное + существительное, Твор.падеж];
[числительное + существительное, Род. падеж + числительное];
[существительное, Им.падеж].

Бұл шаблондарда көрсетілген сипаттамалар, өңделетін құжатқа ең бастапқыда жүргізілген морфологиялық сипаттамалардан алынады.

Төменде осы алгоритм нәтижесінде алынған кілттік сөздер / сөз тіркестері бөлігі көрсетілген:

электронный бизнес, электронная коммерция, данные клиента, высокопроизводительная система, финансовая информация, неоплаченные заказы, первый взнос, конкретный платеж, расчетный банк, другой перевод, сформированный заказ, сеть, графические элементы, курьерская доставка, цифровая подпись, гипертекстовая связь, коммуникация, бюджет, компьютерную сеть и т.д.

Пайдаланылған әдебиеттер тізімі

1. Bellaachia A., AlDhelaan M. HGRANK: A Hypergraphbased Keyphrase Extraction for Short Documents in Dynamic Genre. // Making Sense of Microposts (# Microposts2014)/http://ceur-ws.org/Vol-1141/paper_06.pdf (available: 01.06.2016)

2. Пастушков И.С., Барахнин В.Б. Алгоритм автоматизированного наполнения онтологии фактографической поисковой системы//Всероссийская конференция с международным участием "Знания - Онтологии - Теории" (ЗОНТ-2015), с.87-95, Новосибирск, 6-8 октября 2015 г.

3. Барахнин В. Б., Федотов А.М., Шокин Ю.И. Технология создания программных систем информационного обеспечения научной деятельности, работающих со слабоструктурированными документами // Вычислительные технологии. – 2010 – Т. 15. – N 6.

4. Bougouin A., Boudin F., Daille B. Topicrank. Graph-based topic ranking for keyphrase extraction. In Proceedings of the Sixth IJCNLP, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. – P. 543–551.

5. Gao Y., Liu J., and Ma P. The hot keyphrase extraction based on tf*pdf. In The 2011 IEEE 10th TrustCom, pages 1524-1528, 2011.

6. Hulth A. Improved automatic keyword extraction given more linguistic knowledge. In M. Collins and M. Steedman, editors, Proceedings of the 2003 EMNLP, 2003. – P. 216-223.

7. Liu Z., Huang W., Zheng Y., and Sun M. Automatic keyphrase extraction via topic decomposition. In Proceedings of the 2010 EMNLP, pages 366–376. Association for Computational Linguistics, October 2010.

8. Kushal S. Dave, Varma V. Pattern based keyword extraction for contextual advertising. In CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1885-18887 New York, USA, 2010.

9. Бритков В.Б., Булычев А.В. Методы анализа больших объемов слабоструктурированной информации. – 2010.

10. Шереметьева С.О., Осминин П.Г. Методы и модели автоматического извлечения ключевых слов. – Вестник ЮжноУральского государственного университета, – №1/12. – 2015.