

Питання побудови та використання багатокритеріальної моделі вибору джерел інформації

На цей час користувачу, в першу чергу аналітику у будь-якій сфері - виробничій, економічній, політичній, соціологічній та інших, потрібна інформація, джерела знаходження або походження якої він не в змозі визначити однозначно. Тобто, мова йде про необхідність пошуку та отримання потрібної інформації у безлічі інформаційних джерел [1], які не визначаються жорсткими критеріями. Таким чином, виникає необхідність знаходження у інформаційному просторі тих джерел, які необхідні користувачу, виходячи з багатьох критеріїв, серед яких можна виділити такі найважливіші:

- мінімальна кількість джерел, у значній мірі покриваючих інформаційні потреби. Ця кількість джерел, у кінцевому рахунку, визначає потужність системи моніторингу, яка встановлюється для рішення завдання ефективного інформаційного забезпечення;
- рівень дублювання інформації джерелом. З одного боку, використовуючи відносно повний набір оригінальних джерел, користувач системи моніторингу може відмовитися від джерел з високим рівнем дублювання, але, з іншого боку, їх використання може виявитися достатнім для щоденної неглибокої проробки інформації, тобто відношення до такого роду джерел є неоднозначним;
- тематична стабільність джерел. Природно, що чим вища стабільність та більш вузька тематична спрямованість документів, тим більше спеціалізоване джерело інформації. Дослідження та використання тематично нестабільних джерел є проблематичним;
- тематичний взаємозв'язок джерел, урахування якого дозволяє вибирати тематично корелюючи джерела, виявляти найбільш авторитетні з них (наприклад, центроїди тематичних кластерів) тощо.

Для підтвердження реальної можливості побудови та реалізації в конкретних виробничих умовах багатокритеріальної моделі вибору джерел інформації була проведена дослідна робота щодо визначення необхідного переліку джерел при сформульованих інформаційних потребах аналітиків Національного центру з питань євроатлантичної інтеграції України [2], а саме, отримані документи, релевантні запиту:

(Київ|Київ|Україн|Украин)&(нато|Альянс|євроатлант|євроатлант)

з ретроспективної бази даних системи InfoStream [3] за один місяць (листопад) 2005 року з понад 1520 активних російських та українських джерел інформації в Інтернет (загалом контрольний інформаційний масив складав 1948 МБ даних, 759442 документа). Всі джерела ранжувалися за кількістю їх входжень до результатів пошуку. Результати ранжування дозволили виявити близькість до закону Бредфорда [4], який стосується "паперових" періодичних видань. Джерела розподілилися за трьома рівними за кількістю релевантних документів множинами R1, R2, R3. При цьому R1 – це найбільш рейтингові джерела, що на пряму відносяться до визначеної тематики, R2 – множина джерел, які кореспондуються з тематикою євроатлантичної інтеграції, а R3 – джерела, що лише частково стосуються даної тематики.

З переліку 1521 джерела інформації 615 джерел відповідали запиту. За цим запитом було відібрано 49,5 МБ даних, або 0,025% від вихідного обсягу та 7947 документів (1% від вихідної кількості). У Табл. 1 наведено найбільш рейтингові джерела серед потоків документів, відібраних за запитом. Якщо прийняти позначення, що $|A|$ - це кількість елементів множини A , то пропорція Бредфорда уточнюється таким чином:

$$|R1| : |R3| : |R3| = n : n^{1,2} : n^{2,4}.$$

У даному випадку значення n – складає приблизно 20 джерел. До множини $R1$ можна віднести:

Табл. 1. Найбільш рейтингові джерела для запиту

Ранг	Джерело	Кількість документів у запитах користувачів
1	Євроатлантична Україна	256
2	Український вибір	227
3	ForUM	155
4	Без цензури	136
5	Оглядач	119
6	From-UA	113
7	Обозреватель	113
8	VLASTI.NET	110
9	“Главред”	104
10	“Львівська газета”	102
11	“4 влада”	102
12	Газета “Україна молода”	101
13	Рамблер	94
14	“Диалог.UA”	81
15	“Поступ”	79
16	President.org.ua	78
17	ІАЦ “Ліга”	77
18	АМІ “Новости-Україна”	76
19	“Про новости”	75
20	“Политикум.Инфо”	74

На Рис. 1. наведено рангову діаграму розподілу джерел, що відповідають релевантним запиту документам щодо євроатлантичної інтеграції України.

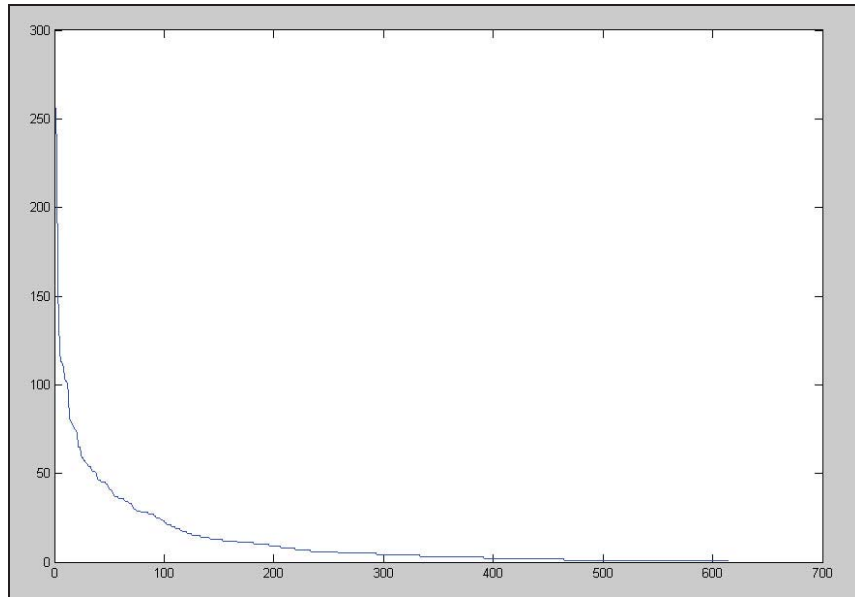


Рис. 1. Рангова діаграма розподілу джерел у множині релевантних документів

На Рис. 2. наведено діаграму розподілу джерел у релевантних запиту документах щодо євроатлантичної інтеграції України у розрізі днів листопада 2005 року. Слід зауважити на значне зменшення публікацій у вихідні дні (субота, неділя).

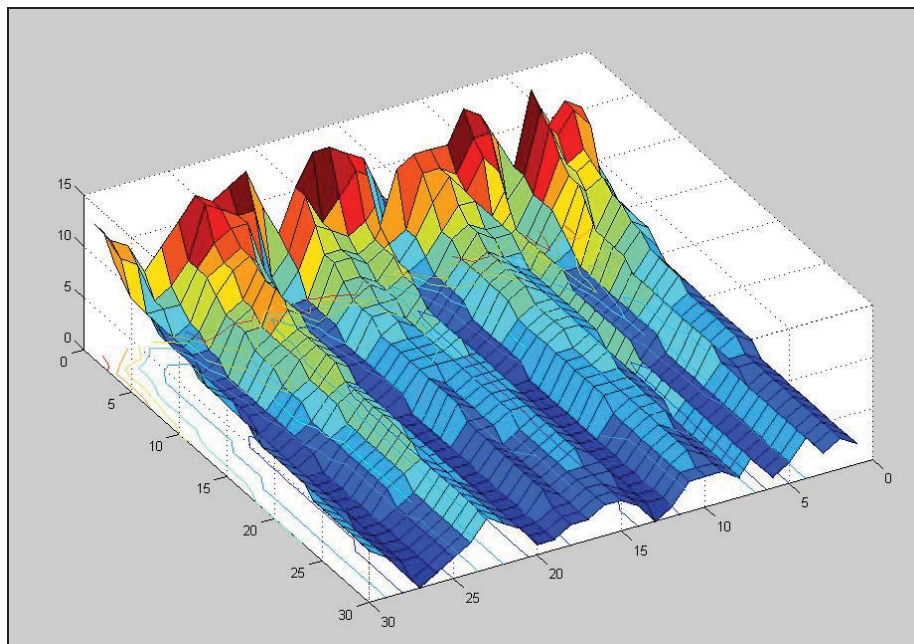


Рис. 2. Рангова діаграма розподілу джерел у множині релевантних документів у розрізі дат

Але при такому підході не виключена можливість надання користувачу значної кількості дублюючої інформації, яку він повинен буде самостійно опрацювати та виключити.

Для виявлення кількості релевантних документів, але таких, що явно дублювалися у системі, було використано пошук за запитом, доповненим логічним виразом "&this.is.double". Оцінка обсягів документів, що відповідали уточненим

запитам, дозволила констатувати, що дублювання складає понад 92%. Серед джерел, що найчастіше використовують дублюючі інформаційні матеріали, можна віднести:

- “VLASTI.NET”;
- “Український вибір”;
- “Євроатлантична Україна”;
- “4 влада”;
- “Оглядач”.

Ці джерела не рекомендується використовувати як базові при пошуку з урахуванням дублювання, але можна розглядати як своєрідні інтегратори новинної інформації [5].

Необхідно враховувати, що велике значення для користувачів має тематична стабільність документів. Рівень тематичної стабільності документів визначається кореляцією наборів тематичних рубрик, які відповідають документам джерел за певний період часу.

Визначення тематичної стабільності документів проводилося за таким алгоритмом:

1. Проводився пошук документів у загальному інформаційному масиві системи моніторингу за визначений період.
2. Формувалася база даних, яка включала код джерела інформації, коди тематичних рубрик, дату публікації документа та унікальні номери документів.
3. Для кожного джерела розраховувався нормований показник тематичної стабільності R за формулою:

$$R = \frac{1}{N} \sum_{i=1}^N \frac{1}{M \cdot \max(\text{rub}_i)} \sum_{j=1}^M \left| \text{rub}_{ij} - \frac{\sum_{k=1}^M \text{rub}_{ik}}{M} \right|;$$

де N – кількість рубрик, M – кількість днів, $\max(\text{rub}_i)$ – нормована максимальна кількість входження рубрики i у документ за весь час, rub_{ij} – нормована кількість входження рубрики i за день j . Процедура обчислення показника для конкретного джерела за визначеною рубрикою мовою програмування Perl така:

```
#---- Обчислення нормованого показника за однією рубрикою
sub sred {
    # @x – масив значень за днями
    $sred=0; # нормований показник
    $max=0; # максимальне значення нормованої рубрики
    for($i=0; $i<$#x+1; $i++) { # цикл за кількістю днів, $i – номер дня
        $sred=$sred+$x[$i]; # накопичення значень рубрик
        if ($max<$x[$i]) {$max=$x[$i];}
    }
    $sred=$sred/($#x+1); # середнє значення рубрик за всі дні
    for($i=0; $i<$#x+1; $i++) {
        $x[$i]=abs($x[$i]-$sred); # обчислення відхилень від середнього
    }
    $sred=0;
    for($i=0; $i<$#x+1; $i++) {
        $sred=$sred+$x[$i]; # накопичення відхилень від середнього
    }
    $sred=$sred/($max*(#x+1)); # нормування відхилень від середнього
}
```

4. Інформаційні джерела ранжувалися за розрахованим показником тематичної стабільності, будувалася відповідна діаграма.

Було проведено дослідження тематичної стабільності (за загальними рубриками) інформаційних джерел, що відповідають запиту. Рангову діаграму розподілу 460 з 1521 джерела, що виходили регулярно протягом листопаду 2005 року, за тематичною стабільністю наведено на Рис. 3. Найбільш тематично стабільними виявилися такі джерела, як “Євроатлантична Україна”, ІАЦ “Ліга”, “Український вибір”.

Результати досліджень, проведених у цьому та попередньому пунктах, дозволяють скласти перелік найбільш тематично стабільних та найбільш оригінальних джерел інформації. Наведені у Табл. 1 джерела інформації охоплюють на цей час приблизно 30% від результатів, що отримані при пошуку за визначеним запитом. Тому як результат цього дослідження, можна рекомендувати включити їх як “Персональні інформаційні джерела” у інтерфейс системи моніторингу Інтернет-ресурсів щодо проблематики євроатлантичної інтеграції України.

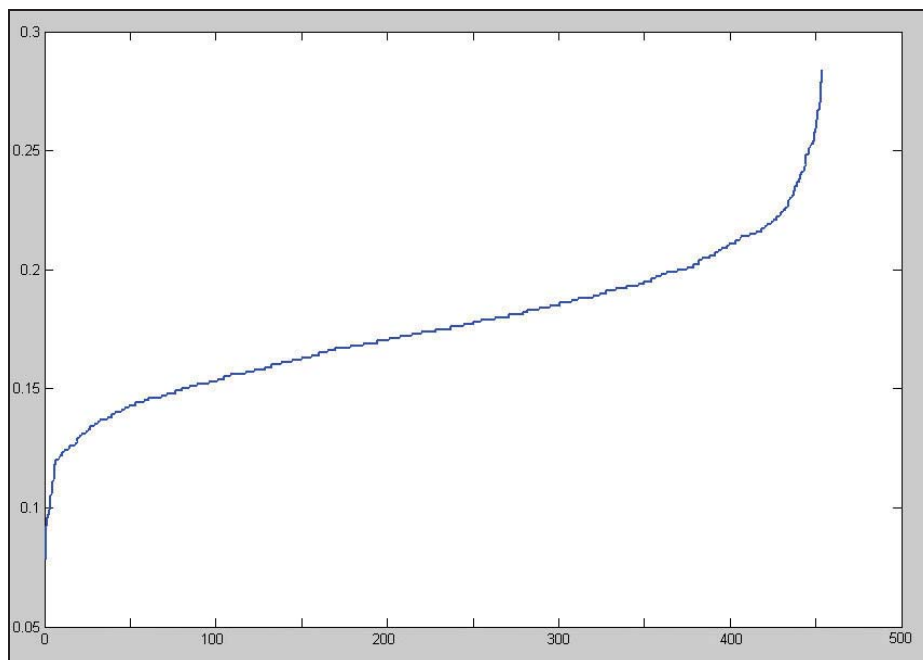


Рис. 3. Діаграма “Ранг документа - Коефіцієнт стабільності”

При аналізі всього масиву інформаційних джерел велике значення має урахування їх тематичного зв'язку. У рамках цієї НДР було побудовано таблицю тематичного взаємозв'язку джерел (ТВД) за таким алгоритмом:

1. Шляхом послідовного сканування текстового масиву документів за визначений період формується таблиця “Джерело – Тематична рубрика”.
2. Розраховується взаємозв'язок джерел за кількістю співпадаючих рубрик.
3. Формується таблиця взаємозв'язку джерел. Джерела з мінімальними зв'язками з іншими (меншими за 5) не входять до розгляду.
4. Методом k-means (k-середнє) здійснюється автоматичне перегрупування таблиці взаємозв'язку джерел (у даному випадку приймається $k=2$).
5. Здійснюється сортування визначених кластерів за кількістю тематичних зв'язків.
6. Кластери об'єднуються, таблиця взаємозв'язку джерел у чисельному вигляді

виводиться в інтерфейсі користувача.

7. Для візуалізації таблиці взаємозв'язку джерел підключаються графічні пакети програм.

На Рис. 4 зображено трьохвимірне представлення таблиці взаємозв'язків джерел, фрагмент переліку яких наведено у Табл. 2.

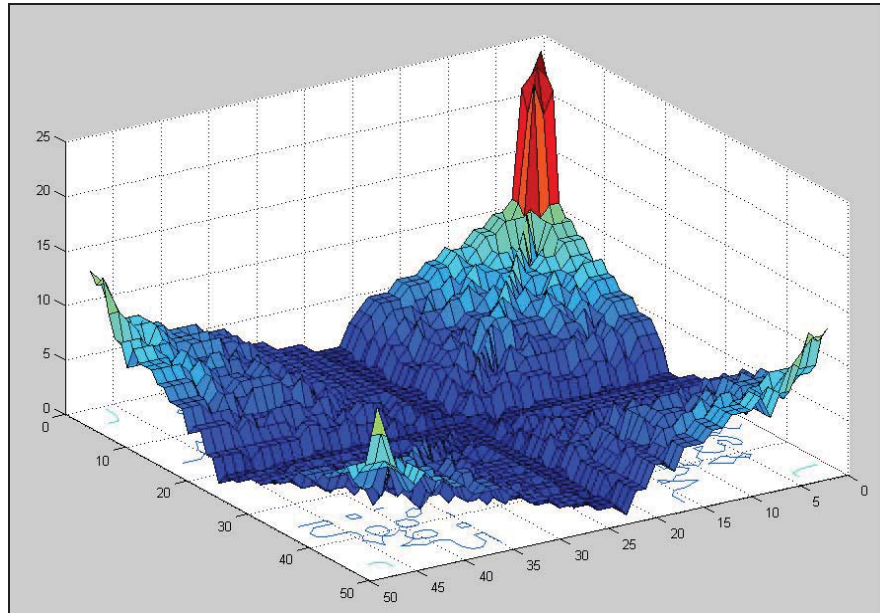


Рис.4. Взаємозв'язок джерел, документи з яких релевантні запиту

Табл. 2. Фрагмент переліку джерел, які застосовувалися у ТВД

№ з/п	№ кластеру	Джерело	Код джерела
1	1	"Обозреватель"	srd00253
2	1	"forUm"	srd00147
3	1	"UA TODAY"	srd00024
		...	
46	2	"Українська правда"	srd00094
47	2	"Без цензури"	srd01424
48	2	"Український вибір"	srd01013

Враховуючи загальний обсяг інформації за період, що досліджувався, а також обсяги інформації, що відповідають запиту за тематикою євроатлантичної інтеграції України можна зазначити:

- загальна кількість джерел, які використовувалися, становить 615, що складає 30% від загальної кількості джерел, що підключено до комплексу моніторингу Інтернет-ресурсів;
- загальний обсяг інформації, яка відповідає потребам Замовника на цей час, становить за зазначений період 49,5 МБ, що складає 0,025% від всієї бази даних, що сканується;
- кількість рейтингових стабільних та оригінальних джерел (всього 20) перебиває 30% інформації.

За місяць об'єм інформації, що передається від підсистеми збору новинної інформації, становить приблизно 2ГБ. Тобто, 0,025% з цієї інформації, або близько

50 МБ на місяць, вирішують сучасні потреби в новинній інформації щодо євроатлантичної інтеграції України. Ці розрахунки дають основу для оптимізації інтернет-трафіку, що споживається.

При виборі джерел для задоволення інформаційних потреб щодо визначеної проблематики у аналітиків завжди виникає питання: “обмежитися вибором невеликої кількості інформативних джерел, або охопити максимальну кількість джерел, і приклавши достатньо серйозні зусилля, виключаючи дублювання та шум, знаходити необхідну інформацію у повному обсязі?” На першому етапі, як правило, приймається рішення за принципом “краще менше, але краще”. Однак, як з’ясовується, такий підхід оптимальний лише для невеликих аналітичних служб. Не охоплюючи максимально повного обсягу інформації, навіть такої, що дублюється, неможливо виділити найчастіше згадувані теми, тенденції, факти. Уникаючи ризику отримання шумової інформації, можна пропустити найбільш свіжі, актуальні повідомлення. Тому, починаючи з певного рівня, аналітичні служби для вирішення завдань інформаційного забезпечення використовують максимально можливий обсяг інформаційних джерел, застосовуючи у своїй роботі розвинені засоби контент-моніторингу.

Наведені вище розрахунки не заперечують концептуальним положенням, що роблять необхідним підключення та використання повного обсягу бази даних:

- користувачі корпоративної мережі Національного центру з питань євроатлантичної інтеграції України повинні задовольняти не тільки типові інформаційні потреби, у полі зору керівництва та фахівців можуть з’являтися нові об’єкти спостереження;
- не завжди текстове представлення інформації, притаманне системі моніторингу, задовольняє потреби користувачів, все частіше необхідно мати доступ до мультимедіа-інформації, для чого потрібно мати безпосередній вихід у мережу Інтернет;
- час від часу виникає потреба використання інформації з закритих ресурсів мережі Інтернет, що на цей час не дозволяє робити функціонуюча система моніторингу Інтернет-ресурсів.

Таким чином, запропонований методологічний підхід щодо побудови та конкретного використання багатокритеріальної моделі вибору джерел інформації доказав своє право на існування та може бути застосований для попередньої оцінки у будь-якої сфері життєдіяльності суспільства.

Список літератури

1. Современные информационные потоки: актуальная проблематика/ Д.В. Ландэ, С.М. Брайчевский // Научно-техническая информация. Сер. 1, 2005. - № 11. - С. 21-33
2. Фурашев В.М., Ландэ Д.В., Григор'ев О.М., Фурашев О.В. Електронне інформаційне суспільство України: погляд у сьогодення і майбутнє: Монографія. – К.: Інжиніринг, 2005. – 164 с.
3. Сканер системы контент-мониторинга InfoStream / Д.В. Ландэ // Открытые информационные и компьютерные интегрированные технологии: Сб. науч. трудов. – Харьков: НАКУ «ХАИ», 2005. – Вып. 28 - С. 53-58
4. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа - М.: "Вильямс", 2005. - 272 с.
5. Определение тематической направленности запросов путем анализа наборов рейтинговых источников / Д.В. Ландэ, С.М. Брайчевский // Открытые информационные и компьютерные интегрированные технологии: Сб. науч. трудов. – Харьков: НАКУ «ХАИ», 2005. – Вып. 29 - С. 41-48.

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ
НАЦИОНАЛЬНЫЙ АЭРОКОСМИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Е. ЖУКОВСКОГО
"ХАРЬКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ"**

**ОТКРЫТЫЕ ИНФОРМАЦИОННЫЕ
И КОМПЬЮТЕРНЫЕ ИНТЕГРИРОВАННЫЕ
ТЕХНОЛОГИИ**

Сборник научных трудов

Выпуск 30

**Харьков
«ХАИ»
2006**

ББК 32.973.3
О 83
УДК 629.7.01; 681.327.11

Утверждено к печати ученым советом Национального аэрокосмического университета им. Н. Е. Жуковского «ХАИ», протокол № 7 от 22.03.2006 г.

Редакционная коллегия: А. И. Бабушкин, П. В. Балабуев, В. К. Борисевич, В. Е. Гайдачук, А. Г. Гребеников, В. В. Кириченко, Б. М. Конорев, В. С. Кривцов, А. К. Мялица, О.И. Парфенюк, В. И. Рябков, И. Б. Сироджа, В. Г. Сухоребрий, О. Е. Федорович, В. Н. Фурашев

Открытые информационные и компьютерные интегрированные технологии: Сб. науч. трудов. Вып. 30. - Харьков: Нац. аэрокосмический ун-т "ХАИ", 2006. 222 с.

ISBN 966-662-129-0

Освещены вопросы аэродинамики, прочности, конструкции и технологии производства авиационной техники, методы интегрированного проектирования, подготовки производства, инженерного анализа с помощью систем CAD/CAM/CAE. Представлены методы обеспечения безопасности конструкций при длительной эксплуатации, системы информационного обеспечения жизненного цикла самолетов, системы обеспечения безопасности жизнедеятельности для летательных аппаратов, вопросы экономического анализа рисков, методы поддержки, принятие управленческих решений

Для специалистов промышленных и научных организаций, преподавателей и студентов.

**Адрес редакционной коллегии: 61070, г. Харьков, ул. Чкалова, 17,
Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ»,
тел. (057) 707-43-63**

**Свидетельство о государственной регистрации КВ №7975 от 09.10.2003 г.
Журнал печатает статьи на украинском и русском языках**

ISBN 966-662-129-0

**© Национальный аэрокосмический университет
им. Н.Е. Жуковского «ХАИ», 2006**

Содержание

П. О. Науменко. О технологическом контроле при первичном формообразовании в современном машиностроении	5
А. Г. Гребеников, А. З. Двейрин, Ю. Н. Геремес, А. М. Гуменный. Метод интегрированного проектирования и компьютерного моделирования фюзеляжа гражданского самолета с помощью интегрированных систем CAD/CAM/CAE/PLM	10
Е. Н. Бут. Технические средства обучения человека-оператора техногенного объекта как предмет изучения эргономики	31
Л. А. Малашенко. Влияние случайных факторов на проектные параметры элементов тонкостенных конструкций	36
В. Г. Сухоребрый, М. И. Никифорова. Оценка влияния корреляции случайных параметров на вероятность потери устойчивости ракеты-носителя	42
И. Б. Сироджа, С. В. Россоха. Нечеткий дедуктивный вывод в системе квантов знаний для поддержки принятия решений при планировании учебного процесса	50
И. Б. Сироджа, И. Е. Россоха. Модель обучения нечеткой базы квантов знаний для поддержки принятия решений в условиях ф-неопределенности	56
А. Н. Нарожный. Компьютерная поддержка процесса определения летно-технических характеристик самолета. Часть 2	62
Д. В. Ланде, В. М. Фурашев. Питання побудови та використання багатокритеріальної моделі вибору джерел інформації	73
А. Г. Гребеников, В. Н. Клименко. Сопротивление усталости полосы с отверстием из титанового сплава ВТ6 в условиях фреттинг-коррозии	80
С. Д. Винничук, В. Ф. Шмырев. Анализ расходов воздуха в раздаточных отверстиях распределительной трубы ВТПОС самолета с учетом докритических режимов течения	90
Д. А. Бастеев. Синтез онтологии управления проектом технологической подготовки производства на машиностроительном предприятии	97
И. В. Бычков. Информационное обеспечение машиностроительного производства	107
Б. В. Лупкин, О. В. Мамлюк, Р. П. Родин. Заточка сверл по цилиндрическим фасонным поверхностям	112
О. В. Мамлюк. Обработка поверхностей типа «переменная малка» торцовыми фрезами	117
В. Г. Сухоребрый, Д. В. Джулгаков. Принципы генерации текстовых эквивалентов для программ, созданных на графическом языке FBD	121
А. В. Ксендзук, И. А. Евсеев, Б. Н. Кухарчук. Исследование влияния нестабильностей на работу бистатической РСА	130