

АВТОМАТИЗИРАНО АНОТИРАНЕ И РЕФЕРИРАНЕ НА ИНФОРМАЦИОННИ ИЗТОЧНИЦИ

Увод

В условията на развитие на информационното общество основна роля играе ефективно действащата система на комуникациите, в която анотациите изпълняват изключително значими функции като посредник между потребителите и първичните и вторичните документи. В огромния поток от документи потребителят на информация се нуждае от определени ориентири за намиране на необходимата му информация. Такъв ориентир е именно анотацията на документите. Анотацията е средство, с чиято помощ потребителят на информация се ориентира в огромния документален поток при избора на интересуващата го информация¹. Тя представлява кратка характеристика на документа или на негова част от гледна точка на съдържание, предназначение, форма и други особености. Анотирането заема важно място в библиотечната, библиографската и информационната дейност за усъвършенстване на информационното обслужване, повишаване оперативността и избирателността на информацията.

Анотацията се съставя за всички видове документи, като широко се използва в каталозите и картотеките на библиотеките, в информационните и книготърговските центрове, служби и фирми. Също така широко се използва в библиотечното, библиографското и информационното обслужване, вкл. и чрез създаваните анотирани БД. Издаваните анотирани библиографии и библиографски указатели помагат на читателите да се ориентират в наличната литература по различни въпроси, облекчават работата на учащите се, преподавателите, учените и специалистите.

Проблемите на анотирането и реферирането са предизвикват повишен интерес у изследователите в областта на библиотекознанието, библиографската, информационната и книготърговската дейност.

Изложение

1. Автоматизирано анотиране на информационни източници.

Понятието „анотация” произлиза от латинската дума „annotatio”, което означава „забележка”, „бележка”, „отмятане”. То е усвоено от много съвременни езици - английски, немски и френски – „annotation”, руски – „аннотация”, полски „adnotacja” и

¹ Гълъбова, С., И. Арсенова. Анотиране на документи. София, Софтрейд, 2004, 3 с.

др. Анотирането на френски език още се среща и като „analyse documentaire”, а на английски - „indicative abstract” (индикативен реферат)². Стандартизирано определение на понятието „анотация” е посочено в Българския държавен стандарт БДС (СТ на СИБ) 2011 - 79 „Реферат и анотация” и гласи: анотацията е кратка характеристика на документа (на част или на съвкупност от документи) от гледна точка на неговото съдържание, предназначение, форма и други особености³. От тази дефиниция следва, че анотацията трябва да дава характеристика на вида на документа, на основната му тема и предназначение, проблемите, обекта, целта на работата и нейните резултати. В нея трябва да се отбелязва и какво е новото, което се съдържа в дадения документ в сравнение с други документи, родствени по тематика и целево предназначение. Необходимо е да се характеризират и особеностите (например при преиздавани документи трябва да се посочва различното между сегашното и предходните издания и т.н.). Според Sayers анотацията може да се публикува заедно с аотириания текст или отделно от него⁴. Когато е отделена от основния текст, анотацията на всяка цена трябва да е съпроводена от библиографско описание на аотириания документ или от други сведения, позволяващи точно да се идентифицира документът.

Задачите на анотацията са да се рекламира, препоръчва и пропагандира четенето на даден документ. Предназначението на анотацията е да даде на читателите само предварителна представа за документа, като им дава точна характеристика на документа, разкриваща неговите силни страни и читателския адрес. Затова тя може да се сравни с портрет на документ, отразяващ неговия образ чрез разкриване на най-важните детайли. Целите на анотацията могат да се групират по следния начин:

1. *Сигнална* – чрез която се сигнализира за появата на даден документ в книжното пространство;
2. *Търсеща* – с която поредица формални и документографски признаци се определя мястото на документа;
3. *Комуникативна* – чрез която се осъществява запознаване със съдържанието или особените белези на документа;
4. *Оценъчна* – с която се прави оценка или препоръка на стила на изложението, на езика, новото, различното, ценното в даден документ.

² Гълъбова, С., И. Арсенова. Анотиране на документи. София, Софтрейд, 2004, 10 с.

³ БДС /СТ на СИБ/ 2011-79 „Реферат и анотация”.

⁴ Sayers, W. First steps in annotation in catalogues. Greenlaw, 1998, 12 p.

В зависимост от целите и задачите съставяната анотация може да бъде справочна, препоръчителна, групова, кратка, разширена и т.н.⁵

Според начина на съставяне анотациите биват обикновени (съставени от човек) и автоматизирани (съставени алгоритмически с помощта на ЕИМ). Автоматизираното аотиране е сложен проблем, чието решаване се базира върху използването на специфични методи за формализиране на текста⁶.

Съдържателите на каналите за оповестяване на различна информация, средства за разпространяване новини, започват да изпитват засилен интерес към процесите на автоматизираното аотиране. Често не само изискванията да се даде възможност за бърз преглед на съдържанието, но и желанието да не се нарушат авторските права са причина пълните текстове на документа да се заменят с техните анотации. Целта е потребителят да може бързо да прегледа големи документи, изчитайки техните кратки анотации.

Всички известни системи за машинно аотиране се основават на принципите на екстрахирането, т.е. на извличането на информация от текста. Програмата не преразказва текста, а извлича онези фрагменти, които смята за важни, и ги обединява в анотация.

Същността на автоматизираното аотиране се състои във формиране на крайно описание на основните теми на текста. Съществуват два подхода за автоматизирано аотиране. В първия случай се извлича малък брой изречения, съществуващи в текста, които най-пълно отразяват неговите основни теми и микротеми. В тях трябва да се съдържат и ключови думи. Ако липсват, те се извличат от текста. Във втория подход от текста се извличат мисли. Тези мисли се подреждат в изречения, които съставят нов текст, наречен анотация. Вторият вариант е значително по-сложен. Например SomeP технологията позволява автоматичен синтез на кратки, прости фрази и изречения.

Като цяло задачата на автоматизираното аотиране е да се определи тематиката на документа, да се извлекат ключови думи и изрази с оглед на смисъла, да се търсят изрази, съдържащи ключови думи и фрази, както и изречения, отразяващи основните теми на текста. Най-развитите средства за аотиране отчитат зависимостта на изреченията едно друго, за да се получи свързан текст, като програма подбира групи взаимосвързани изречения, а след това „ги залепва“ едно за друго. Все още слабо звено

⁵ Гълъбова, С., И. Арсенова. Аотиране на документи. София, Софтрейд, 2004, 22 с.

⁶ Пак там, 43 с.

в автоматизираното аотиране са синтактичният разбор и използването на тезаурос (речник на термини в определена област).

Автоматизацията на процесите на аотиране в България се движи в унисон със световните тенденции. В областта на програмите за информационно общество Европейският съюз разработва проект, първата част от който е посветена на автоматизираното аотиране на документи. Тази част е фокусирана върху метода, използван за описване и аотиране на електронната информация. Чрез използване на семантични техники потребителят ще може да ограничи резултатите до темата, която търси в момента. Консорциум от водещи европейски университети разработва нови методи за аотиране на информация, които се основават на управлението на метайнформацията.

Софийският университет е разработил автоматична информационна система „Архив”, предназначена за съхраняване и обработване на архивни документи с неограничена тематика и предоставяне на потребителите на документи от архива по техни характеристики (обекти, тематика, ключови думи и др.) Информацията може да бъде въвеждана и извеждана в интерактивен режим, чрез сканиране на печатни материали и от текстови масиви. При въвеждането на информация се осъществява формален и логически контрол на данните, за да се осигури коректната им обработка. Електронната система библиотека се разработва с цел новите книги да бъдат по най-бърз начин представяни с помощта на автоматизираното аотиране. Тя е създадена за работа в среда на операционните системи Unix, Windows, а системата за управление на бази данни е поверена на релационна система Oracle. Въвеждането на данните за изданието включва аотиране. Според тяхното класифициране и потребителски номенклатури. Във Физическия факултет на СУ е създадена програма за автоматично аотиране „Annotate”, с която всички документи във факултета се представят с техните кратки аотации. От началото на 2001г. немската фирма „Немечек” предоставя на България системата за управление на електронни документи „DocuWare”, ориентирана към средни и големи фирми. Софтуерът осигурява въвеждане, сканиране, конвертиране, индексирание, извличане на информация, отпечатване и издаване на най – разнообразни документи.

Автоматизираното аотиране може да бъде използвано на различни места - например в различни видове библиографски издания, картотеки и каталози (вкл. в библиотечни и книготърговски), в книги, списания и др., в библиографски справки, списъци и други библиотечни и информационни продукти. Съдържанието на

автоматизирани анотации е в пряка зависимост от целта, за която се съставят, затова те трябва да са такива, че да осигуряват осъществяването ѝ.

2. Автоматизирано рефериране на информационни източници.

Реферирането се смята за сложна интелектуална дейност, насочена към смислово преобразуване на съдържателната информация на текста и за това тя се обработва от системите за семантичен анализ на информацията. Под автоматизирано рефериране се разбира машинното преработване на машинни текстове, при което от тях се отделят изречения, отразяващи основния смисъл на съдържанието. Този набор от изречения не е истински реферат, затова се говори за „квазирефериране”. При автоматизираното рефериране целта е да се разкрият основните и инвариантните елементи от съдържанието на текста за експлициране на основното съдържание на документите.

Работата на системите за автоматизирано рефериране се основава на формализация на процесите при семантичното свиване на информация, като се отчитат процесите на декодиране на семантиката на първичния документ. В научните текстове семантиката изобразява някакви фрагменти от системата на знанията, където присъстват общи устойчиви елементи на ситуациите и основните отношения между тези елементи. Ето защо методите на автоматизираното рефериране се базират на разкриването и отделянето на основните и инвариантните елементи от съдържанието на текста. Автоматизираното рефериране се основава на статистическия анализ на текста и се провежда по различни методики:

1. Оценка на степента на значимост на изречението, която се изразява като отношението на квадрата на броя на значимите думи в изречението към общия брой думи в изречението, където S е оценката на степента на значимост, M е броят значими думи, а N е общият брой на думите в изречението.

2. Оценка на степента на значимост на изречението, изчислена като най-голям брой срещани се в него еднакви значими думи, като се отчита честотата на употребата на отделни значими думи.

3. Оценка на значимостта на изречението, изразена като честота на употребяването на влизащите в него значими думи.

Познати са няколко метода за автоматизирано производство на индикативни реферати:

1. Метод, използващ принципа на местонахождението, който предполага, че информацията в изречението може да се отнесе към неговото място в статията и се приема, че най-важни са първите изречения на абзаците, особено на уводния.

2. Метод, използващ индикативни словосъчетания като: „целта на статията се състои във”, „целта се определя”, „в заключение може” и т.н.

3. Метод, използващ честотата на ключовите думи.

4. Метод, използващ честотата на използването в изреченията на думи от заглавието на статията, от подзаглавието или рубриците.

Лингвистичните методи се основават на изследване на лингвистиката и семантиката на текста и разработване на специални алгоритми за извличане на сведения с помощта на речник индикатор. Всички съществуващи алгоритми за рефериране се основават на избора на готови изречения от текста, които съдържат най-голям брой често срещани понятия. По такъв начин информационния приоритет на документа се разкрива под формата на последователности от цитати, подбрани от първичния документ. Наличието на мрежа от понятия и свързани глаголи позволява да се формулират основните идеи на текста във вид на прости изречения.

Състоянието на работите по автоматизираното рефериране се характеризира по следния начин:

1. Създаването на машинни реферати се основава на извличане на отделни изречения от текста на първичния документ.

2. Автоматизираното рефериране се основава на статистически критерии, разкриване на ключови думи, наличието на определени елементи в изречението, разположени в началото или в края на параграф или раздел, логически връзки на изреченията в текста.

Всяка система за автоматизирани реферати включва два задължителни компонента:

1) Word Control List – списък от контролни думи, съставен с текстове на определена тематика;

2) набор от правила за обединяване на думите от списъка във фрази за построяване на изречения. В процеса на рефериране компютърът автоматично забелязва определени думи, фрази и изречения. Останалият текст представлява екстракт, наречен „машинен реферат”. Операциите за съставянето на такъв реферат са: 1) четене на документ; 2) анализ; 3) прилагане на правилата за набор от думи; 4) съставяне на реферат; 5) възпроизвеждане на реферата във вид на текст.

Цялата съвкупност от методи, използвани за автоматизирано рефериране може да се раздели на статистически логико-математически и лингвистични методи (Леонов 1975). Съществуващите методи за автоматизирано рефериране все още не дават

удовлетворителни резултати. В ранните етапи на работата по автоматизирано рефериране машинните реферати се разглеждат най-вече като илюстрация за интелектуалните възможности на персоналния компютър. Яви се обаче труднопреодолимо препятствие пред автоматизираното рефериране а именно - недостатъчното познаване на структурата на естествения език. За най-перспективен метод се смята лингвистичният, при чието усъвършенстване се цели използването на парефразиран текст и изучаването ролята на контекста в определени информационни концепции.

Реферативните бази данни могат да се разделят по организационно-отраслов признак на следните две групи:

1) *политематични* - обхващат много области на знанието; техни представители са базите данни Винег и Bulletin Signalitique, както и реферативни бази за патенти от различни страни.

2) *тематични (специални)* - обхващат само една област на знанието; представители са Chemical, Biological, Physical Abstracts.

Реферативните бази данни и електронните версии на реферативни издания имат следните предимства: икономия на хартия; по-малка стойност на съхранението; изключват се разходите по подвързване на печатните версии; по-ефективно обслужване и увеличаване фонда на електронните документи; дават възможности за многовариантно обслужване.

За системата на библиотечно-информационното обслужване са характерни няколко типа задачи. Задачата за контекстно-свободно търсене на информация е една от основните. Най-прост е методът „търсене по формален признак” - сходство на речниковия състав на запитването и подбраните документи. Най-качествените системи предполагат разширяване на запитването чрез морфологически словоформи, а също така синоними от предоставен разработен речник. Основа за такъв подход е следното съждение - ако документите си приличат по състава от думи, с голяма степен на вероятност те си приличат и по смислово съдържание. Затова следващата стъпка е да се смени критерият за сходство от „сходство по речников състав” в сходство по други признаци, които биха могли по-точно да отразят сходството на смисловото съдържание.

Заклучение

Още в края на XX в. обществото ни да премина в друг етап от своето развитие, наречен информационен. Информацията стана стратегически ресурс и се разкриха неограничени възможности за нейното използване. Потребителите все повече се стремят да получават възможно най-пълна и точна информация с минимални усилия и разход на време. Това изисква ново осмисляне на аналитико-синтетичното преработване на информацията. В тази връзка проблемите на аотирането и реферирането придобиват актуално звучене.

В съвременната информационна дейност се отделя голямо внимание на различните форми за свиване на информацията, най-разпространената форма на която е рефератът. По такъв начин до потребителя стигат основното съдържание и изводите на документите. Реалностите на днешния ден поставят по различен начин подхода към информацията, определят необходимостта от осмислянето на нейната роля и форма на въздействие. Анотацията и реферирането си остават едни от най-ефективните средства за сигнализиране и търсене на информация в съвременния свят. Това налага да се обърне сериозно внимание и да се търси комплексен подход при разглеждане на процесите на автоматизираното аотиране и рефериране на информационни източници. Анализът на тези процеси показва, че те са е неделима част от библиотечната, библиографската и информационната дейност на съвременното технологично общество.

Библиография:

1. Арсенова, И., С. Гълъбова. Рефериране на документи. София, За буквите – О писмменехъ, 2006, 186 с.
2. БДС /СТ на СИВ/ 2011-79 „Реферат и анотация”.
3. Гълъбова, С. Интегративност на източниците на научна информация, стандартизация, интелектуална собственост. София, WINI, 2009, 412 с.
4. Гълъбова, С., И. Арсенова. Аотиране на документи. София, Софтрейд, 2004, 103 с.
5. Денчев, С. Информационна среда за трансфер на технологии. София, З.Стоянов, 2003, 166 с.
6. Томов, П., А. Ангелов. Специфика на съвременните системи за автоматизиране на превода. (07.09.2014) <www.tu-sofia.bg>.
7. Конгресна библиотека на САЩ. MARC 21 – формат за библиографски данни. (07.09.2014) <www.nalis.bg> .
8. Sayers, W. First steps in annotation in catalogues. Greenlaw, 1998, 12 p.