

קריטריונים לבחירת מנוע אחזור טקסט

גרסה 4 – 7.2005

עפר דרורי

offerd@gmail.com

מבוא

מטרת מסמך לזה להגדיר טבלת קריטריונים לצורך השוואה בין מנועי חיפוש לאחזור טקסט. עם הגידול במאגרי המידע ובכמויות המידע המילולי בהן (בעיקר טקסטים שאינם מפורמטים) עולה הצורך בכלים / מוצרים אשר ינהלו את המידע הרב הזה. ניהול טקסטים חופשיים בצורה איכותית, בעיקר בכל הקשור לאחזור המידע מהם, הוא משימה מחשבתית שאינה קלה כלל ועיקר. בגלל מורכבות המשימה מחד והצורך בפתרון בעל ביצועים סבירים מאידך נוצרה בתחום זה התמחות אשר מסופקת על ידי מספר רב של חברות בעולם. נפוצותם של מערכות לאחזור טקסט ומנועי חיפוש גדלה באופן משמעותי יחד עם התפתחות רשת האינטרנט והצורך לאיתור מידע מהרשת. ככל שהמאגרים גדולים יותר וככל שהמידע המוזרם לתוכם לא עובר תהליכי סינון ובקרה, כך גדלה חשיבותם של המנועים ובמיוחד תכונותיהם. במאגרים מקצועיים בהם הבקרה על המסמכים רבה, הצורך בתכונות משוכללות של המנוע קטנה במידה מסוימת אף שגם במאגרים מסוג זה קטלוג החומר ומפתוחו לא תמיד עונים על כל הצרכים. בנוסף אי אפשר להתעלם שגם במאגרים המקצועיים נעשים ניסיונות "לעיגול פינות" בגלל העלויות הגדולות הכרוכות בבקרה זו.

המסמך כאמור, מרכז קריטריונים רבים לצורך בדיקה והשוואה בין מוצרים שונים והוא מהווה כלי עזר לבחירה של מוצר לארגון על פי הצרכים הארגוניים שנקבעו מראש. הטבלה כוללת התייחסות למספר נושאים:

נתוני זיהוי של המוצר, החברה המפתחת והנציגות בארץ אם קיימת
תכונות המוצר, בקריטריונים השונים
דרישות פיתוח, ההנחה היא שמנוע החיפוש ישתלב במערכת גדולה יותר וחלק מהיכולות שלו יבואו לידי ביטוי במערכת זו באמצעות כלי פיתוח מתאימים
טכנולוגיה של המוצר, יכולת טכנולוגית, תמיכה וקישוריות למוצרים אחרים
ביצועים של המוצר
מידע על הספק וניסיון, נושא חשוב לעיתים אף עולה בחשיבותו על תכונות מסוימות במוצר
מחיר המוצר, בהתייחס לתצורות השונות

מומלץ להעביר את הטבלה המצורפת להתייחסות הספקים השונים. יש להחליט כמובן על השקלול המתאים של התכונות השונות עפ"י מודל דלפי לברירת חלופות.

פורמט Word של המסמך נמצא בכתובת

<http://www.sigtrs.org/a.php?c=sigtrs&a=287&rc=sigtrs>

פורמט PDF של המסמך נמצא בכתובת

<http://www.sigtrs.org/a.php?c=sigtrs&a=287&rc=sigtrs>

מסמך המפרט את ספקי מנועי האחזור התומכים בעברית נמצא בכתובת

<http://www.global-report.net/a.php?c=drori&a=57&rc=drori>

מידע נוסף על תכונות המנועים, אחזור טקסט ועוד נמצא באתר קבוצת העניין אחזור טקסט – SIGTRS

<http://www.sigtrs.org>

פרטים מזהים

	שם המוצר
	שם קודם של המוצר
	מספר גרסה נוכחית
	שם החברה המפתחת
	כתובת החברה
	שם הנציגות בארץ
	כתובת הנציגות
	שם איש הקשר בארץ
	טלפון איש קשר
	פקס איש קשר
	דואר אלקטרוני
	אתר אינטרנט של המוצר

תכונות

התייחסות הספק	הסבר	קריטריון
	תכונה בסיסית, בהתייחס לטקסט חופשי	אחזור על מסמכי טקסט
	היכולת לבצע אחזור על שדות מידע רגילים מפורמטים	אחזור על מידע מפורמט
	כמו תאריך, מחבר וכו'	אחזור לשדות מפורמטים בתוך מסמך טקסט
	,=, <, >, OR, NOT, AND שימוש ב- () לביטויים מורכבים	אופרטורים בולאנים
	"מילה" AND "מילה שנייה" במרחק X מילים, באותו משפט, באותה פסקה וכו'	אופרטורים מטריים (מטריקה)
	הכוונה לטיפול מיוחד בשפה ולא לאחזור על בסיס ייצוג האותיות העבריות במאגר	אחזור לשפה עברית
		אחזור לשפה אנגלית
	באותו מסמך	אחזור ל- 2 השפות במערב
	נא לפרט	אחזור לשפות נוספות
	האם קיים כזה ואם כן מהן תכונותיו	מילון מורפולוגי כחלק מהמוצר
	האם השילוב אפשרי, אם כן ציין איזה מילון (כמו מורפיקס של מלינגו) ואיזה גרסה	שילוב מילון מורפולוגי חיצוני
	האם קיימת תשתית לשימוש בטבלאות תזאורוס חיצוניות, האם המוצר כולל תזאורוס משלו, אם כן לאיזה תחום ובאיזה שפה	תמיכה בטבלאות תזאורוס במוצר
	במידה ואנו רוצים לייצר את המילון לבד, האם יש תמיכה לשימוש בטבלה ריקה שתזוץ ע"י המשתמש	ניהול תזאורוס במוצר
	האם קיימת תשתית במוצר לתזאורוס היררכי	ניהול תזאורוס היררכי
	האם המוצר כולל ממשק משתמש להצגת תוצאות החיפוש, אם כן באיזו שיטה: רק כותרות, תחילת מסמך, משפטים רלוונטיים לחיפוש וכו' – יש לפרט	הצגת תוצאת החיפוש ע"י המוצר
	האם המוצר כולל ממשק משתמש לביצוע שאילתת החיפוש	הצגת שאילתת החיפוש ע"י המוצר
	האם המוצר כולל תמיכה בהדגשת מילים המקיימות את תנאי החיפוש במסמכים שאותרו	הדגשת מילים המקיימים את תנאי החיפוש
	יצירת סטים של תשובות וביצוע אחזור נוסף עליהם	ביצוע שאילתות על שאילתות
	האם אפשרי ואם כן האם גם בפורמטים שונים של המאגרים	אחזור על מספר מאגרים במקביל
	טיפול במצב מורכב בו כותרות המסמכים נמצאות בבסיס נתונים והמסמכים עצמם במערכת לניהול קבצים	אחזור על מסמכים מפוצלים
	האם קיים בלי הרחבות אוטומטיות	אחזור מדויק בלבד
	האם קיים	אחזור פונטי

	ראשיות, סופיות ואמצעיות אפשרות להחלפת תווים בסימן "*" "	הרחבות אחזור לחלקי מילים (Wildcard)
	האם נעשה במוצר עצמו, האם מתבסס על בסיסי נתונים חיצוניים, תיאור שיטת עבודה	ניהול אינדקסים במוצר
	האם יש תמיכה להרשאות שונות על קטעים במאגר	ניהול הרשאות גישה לקטעים במאגר
	האם כולל מנגנון גיבוי ושחזור עצמי או שמתבסס על גיבוי ושחזור חיצוני של כל סביבת העבודה	גיבוי ושחזור במוצר
	האם קיים, מגבלות אם יש נא לציין	תמיכה באחזור מסמכי טקסט ASCII
		תמיכה במסמכי UNICODE Text
		תמיכה במסמכי XML כולל מסמכי אופיס שנשמרו כ- XML
	להתייחס לאיזה גרסאות WORD, הכוונה לאחזור ישיר ממסמכי WORD ללא הסבתם	תמיכה באחזור מסמכי טקסט מסוג WORD
		אחזור מתוך מסמכי RTF
	האם קיים	תמיכה באחזור מסמכי טקסט מסוג HTML
	האם קיים, ציין מגבלות לגבי לוגית וחזותית	תמיכה באחזור מסמכי טקסט HTML בעברית לוגית וחזותית
	האם קיים, הכוונה לאחזור ישיר ממסמכי Excel	אחזור מתוך מסמכי Excel
		אחזור מתוך מסמכי Powerpoint
	האם קיים, ישירות מול קבצי אקרויבט PDF	אחזור מתוך מסמכי PDF
	האם קיים, ישירות מול מסמכי Post Script	אחזור מתוך מסמכי PS
	האם המוצר כולל תרגום פורמטים שונים לטקסט נקי, אם כן פרט אילו	תרגום פורמטים
	האם קיים	אחזור לשורשים בשפה העברית והאנגלית
	מתוך מערכת O.L, הכוונה למנגנון המאפשר הוספה מיידי של מסמך וביצוע אחזור מיידי למסמך במסגרת המוצר ולא להפעלת אצווה כל מספר שניות או דקות	אפשרות להוספת מסמכים בצורה מקוונת
	אחזור וניהול המידע כאשר הוא בטבלאות	תמיכה בטיפול בטבלאות
	האם המוצר כולל מנגנון לדירוג רשימת התוצאות, פרט אלו אלגוריתמים קיימים לדירוג	דרוג (Ranking) תשובות
	האם המוצר כולל מנגנוני סיווג כמו קטגוריזציה, אשכולות דינמיים ועוד. נא לפרט גם בהתייחס לשפה	מנגנונים לסיווג

	<p>יכולת המוצר לתמוך במנגנוני אבטחה (קבצים או בסיסי נתונים) כך שהמשתמש יקבל את רשימת תוצאות החיפוש בהתבסס על ההרשאות שלו</p>	<p>אבטחת מידע</p>
--	--	-------------------

דרישות פיתוח

התייחסות הספק	הסבר	קריטריון
	<p>יכולת המוצר בתמיכה בסביבות פיתוח שונות SDK ל- Java .Net וכו'</p>	<p>תמיכה בסביבות פיתוח</p>
	<p>קיום SDK למגוון צרכים כמו: חיפוש, קלסיפיקציה, תחזוקת אינדקסים, תחזוקה כללית של מנהל המערכת, יכולות חיפוש מורחבות וכו'</p>	<p>קיום SDK</p>

טכנולוגיה

התייחסות הספק	הסבר	קריטריון
	האם קיימת	תמיכה בשרת/לקוח
	התייחס לגרסאות התומכות בסביבות העבודה השונות, לגבי דפדפנים התייחס לגרסאות השונות, אם ידועות בעיות בנושא - נא ציין	תמיכה בלקוח תחת חלונות XP, 2000, 98 וכו' דפדפנים לסוגיהם
	האם קיימת	תמיכה בשרתי NT, UNIX
	האם קיימת תמיכה למידע המאוחסן בבסיסי הנתונים הנייל, אם קיימת תמיכה בבסיסי נתונים מקומיים אחרים נא ציין	קישור לבסיסי נתונים מסוג SQLServer, אורקל
	האם קיים בכל הקשור לאחזור מידע המוטמע בשרתי Exchange	קישור ל- Exchange
	האם קיים בכל הקשור לאחזור מידע המוטמע בשרתי Domino	קישור ל- Notes
	ציין אילו קיימים ואם קיימים נוספים נא ציין	ממשקים מסוג Java, OLE, RPC, ACTIVEX
	האם קיים ממשק API המאפשר ביצוע של כל התכונות מתוך תוכנה אחרת חיצונית	ממשק API
	האם קיים	קישור לשרת אינטרנט IIS
	האם קיים	קישור לשרת אינטרנט של אורקל
	הכוונה למצב בו המידע המיועד לאחזור מאוחסן בבסיס נתונים מרכזי ב- M.F. כדוגמת ADABAS והאינדקסים לאחזור נמצאים בשרת או ב- M.F. (נא ציין במה המוצר תומך)	קישור לבסיס נתונים ב- MF כאשר אינדקסים נמצאים ב- MF או בשרת
	מה המספר המקסימאלי של אינדקסים שניתן להקים ולנהל?	תמיכה בריבוי אינדקסים
	האם מתבצע אינדוקס אוטומטי עם עדכון בסיס הנתונים?	אינדוקס ישירות מבסיס הנתונים
	האם יש אפשרות לכלול את האינדקסים בבסיס הנתונים עצמו?	שמירת האינדקס בבסיס הנתונים
	האם קיימת	תמיכה בבסיס נתונים מעל 10 מיליון מסמכים

ביצועים

התייחסות הספק	הסבר	קריטריון
	אינדוקס של 100,000 רשומות חדשות, נא ציין ביחס לשרת סטנדרטי (ציין חוזק השרת, מספר CPU וכו')	זמן טעינה של 100,000 מסמכים
	כני"ל לגבי 10,000	זמן טעינה של 10,000 מסמכים
	עפ"י מבחני ביצועים של המוצר (נא ציין המבחן)	זמן תגובה לחיפוש ביחס לגודל המאגר
	בזמן נתון ועם גידול המאגר, נא לצרף נתונים	נפח אינדוקס ביחס לגודל המאגר

מידע על הספק וניסיונו

התייחסות הספק	הסבר	קריטריון
		ניסיון בשנים בפיתוח מנוע אחזור טקסט
		מספר התקנות בארץ
		מספר התקנות בחו"ל
		כמות התקנות למאגרים מעל 10 מיליון מסמכים
		כמות התקנות מעל מיליון מסמכים
		כמות התקנות מעל 100,000 מסמכים
		כמות התקנות בסביבת אינטרנט/אינטרה-נט
	נא לציין את סוג בסיס הנתונים: אורקל, SQL-Server וכו'	כמות התקנות המשלבת בסיס נתונים ארגוני על PC
	נא לציין את סוג בסיס הנתונים: אדבס, DB2 וכו'	כמות התקנות המשלבת בסיס נתונים ארגוני על MF
		כמות התקנות עם Exchange
		כמות התקנות עם Notes
	נא לפרט: תמיכה טלפונית או אחרת	סוגי תמיכה
		פרק זמן בין גרסה לגרסה
	האם מוגבלת בזמן, האם מרכז עיסוק החברה בתחום, תוכניות לעתיד בהקשר למוצר	מחויבות החברה למוצר
	עפ"י הקריטריונים השונים: לקוחות עם בסיס נתונים אדבס, DB2 אחר לקוחות עם בסיס נתונים אורקל, SQL-SERVER וכו'.	רשימת לקוחות
		מספר שנים שהמוצר הנוכחי מותקן אצל לקוחות

	השתתפות במכרזים או בחירות קודמות	אם המוצר השתתף במכרזים או בחירה אחרת ונבחר, נא ציין את הגורם הבוחר, שנת הבחירה ואת המוצרים שמולם עמד לתחרות
--	----------------------------------	---

מחיר

התייחסות הספק	הסבר	קריטריון
	נא להתייחס לגרסאות שונות של המוצר (אם קיימות) בהתייחס למספר שרתים, לאתר, שיטות רישוי שונות	מחיר המוצר
	נא לציין את תקופת האחריות	מחיר אחזקה לשנה לאחר תקופת האחריות