

## מנוע חיפוש קטגורי מבקר: אבטיפוס / שמואל גוטמן

עם התפתחות רשת האינטרנט שחלה בשנים האחרונות, איתור מידע ברשת האינטרנט הפך לאחד השימושים העיקריים של המשתמשים ברשת. למרות השיפורים שחלו במנועי החיפוש השונים בתחום היקף המידע והשיפור באלגוריתמי החיפוש והדירוג, עדיין נתקלים מחפשי המידע באשר הם, בכמה בעיות עיקריות הגורמות לכך שתוצאות החיפוש אינם עונות על צרכיהם. לעיתים, מדובר בכך שתוצאות החיפוש כלל אינם קשורות למונח אותו חיפש המשתמש ולעיתים, תוצאות החיפוש אמנם מכילות מילות חיפוש אותם ביקש המשתמש אך בהקשר של נושא אחר לגמרי. כמו כן, מחקרי שימושיות שנעשו, מצביעים על כך שהמשתמש הממוצע בודק רק את התוצאות הראשונות במנוע החיפוש. לכן כאשר מידע רלוונטי אינו מופיע בראש רשימת התוצאות, המשתמש לא נחשף אליו כלל. בעיות אילו גורמות לכך שבפעמים רבות המשתמש מתייאש ממצאת המידע או שאף גרוע מכך ובידי המשתמש נותר מידע חלקי/מוטעה. במקרים רבים אחרים, המשתמש נאלץ לנסות שאילתות שונות וחוזר על התהליך כולו מספר פעמים עד למציאת המידע הרצוי. תהליך שלוקח זמן ועמל רב.

מטרת המחקר היא לתת מענה לבעיות אלו וליצור מערכת חיפוש חדשנית שתתן למשתמש את היכולת להתמקד במהירות בתוצאות חיפוש איכותיות, בתחומים הרלוונטיים אליו. כדי להשיג את המטרות הללו המערכת תעשה שימוש במנוע חיפוש ובמדריך אתרים קיימים ותשלב מידע משניהם כך שתוצאות החיפוש שהתקבלו ממנוע החיפוש, יוצגו בפני המשתמש כשהן מסודרות לפי קבוצות של נושאים על בסיס המידע הקטלוגי הקיים עליהם במדריך החיפוש. בעזרת השימוש במערכת, ניתן יהיה למקד את המשתמש במהירות במידע המבוקש, לפרוס בפניו את המידע שהתקבל בצורה ברורה ולהציג מידע איכותי ומבוקר ובכך להקל על המשתמש.

מערכת זו, נקראה בשם 'מנוע חיפוש קטגורי מבוקר' או בקיצור CSE. זהו ראשי תיבות הן של Category Search Engine והן של Controlled Search Engine. בליבו של CSE מצוי אלגוריתם חדשני. אלגוריתם זה עושה שימוש במנוע חיפוש גוגל לצורך אחזור המידע מרשת האינטרנט ובמדריך האתרים של יאהו לצורך קבלת מידע אודות התוצאה שנתקבלה. האלגוריתם מתשאל את מנוע החיפוש במונח החיפוש המבוקש. ולאחר מכן, האלגוריתם עובר באופן סדרתי על כתובות התוצאות (URL) שהתקבלו ממנוע החיפוש. ומחפש תחת איזה נושא (קטגוריה) מופיעה הכתובת שהתקבלה מגוגל, במדריך האתרים של יאהו. במידה

והכתובת המדויקת לא נמצאה במדריך האתרים האלגוריתם מנסה לאתר כתובת חלקית או לפחות את הנושא אליו שויך האתר בו נמצא המסמך. בדרך זו האלגוריתם מספק עבור כל מונח חיפוש אוסף של תשובות ממנוע החיפוש, כאשר לכל תוצאה ישנו מידע נוסף בדבר הקטגוריה אליו התוצאה שויכה ביאהו. לאחר הפעלתו של האלגוריתם, CSE מציג את תוצאות החיפוש והמידע הקטלוגי שנאסף על ידי האלגוריתם, באופן דומה במקצת, לנעשה במנועי האשכול (clustering) הקיימים. כאשר, כל תוצאה ממוקמת בעץ קטגוריות על סמך המידע שנתקבל ממדריך האתרים. באופן זה CSE רותם, מחד גיסא, את יכולות החיפוש בטקסט ואת ההיקף של מנועי החיפוש, ומאידך גיסא את ההיבט של ההבנה, הבקרה האנושית והחלוקה לנושאים שמצויה במדריכי החיפוש. כמו כן, שיטה זו תנצל את היתרונות שמצויים בשיטת התצוגה של אשכול, שעיקריה הן שיפור יכולת האחזור של המשתמש על ידי מיקוד בנושא הרצוי ללא צורך לנבור בין עשרות תוצאות לא רלוונטיות ופריסה נושאית של תחום העניין המבוקש. יתרונו של CSE על פני מנועי האשכול הקלאסיים הינו בכך שנושאי החלוקה נוצרו מכוח ניתוח טקסטואלי סטטיסטי אלא על סמך חלוקה מבוקרת אנושית.

המחקר מורכב משלושה שלבים:

1. איתור כלים ומשאבים הדרושים לביצוע המחקר.
2. פיתוח וניסוי ראשוני של המערכת לצורך בדיקות היתכנות ושיכלול המערכת.
3. הערכה וניסוי של המערכת מול מנועי חיפוש מתקדמים.

האלגוריתם של CSE עושה שימוש במנוע חיפוש ובמדריך אתרים קיימים. לכן בשלב הראשון היה צורך לאתר מנוע חיפוש ומדריך אתרים שיתאימו לצורכי המחקר. המחקר מכוון למשתמש הממוצע ולכן מנוע החיפוש שנמצא מתאים למחקר, הוא גוגל, שנחשב למנוע החיפוש המקובל ביותר בעולם. מדריך האתרים שנמצא מתאים למחקר, הוא המדריך של יאהו, שהצטיין בגודלו, היקפו, ומבנה היררכי ברור. לצורך הערכה של CSE היה צורך לאתר מנוע אשכול קיים. מנוע האשכול שנבחר הוא carrot2.org. הסיבה לבחירה בו היא בגלל שהוא המנוע היחידי שחושף את האלגוריתם שמאחוריו ופועל ברישיון חופשי. לצורך הניסוי והפיתוח אותרו 50 שאילות פופולאריות שפורסמו בפרסומים סטטיסטיים.

לצורך ניסוי ופיתוח CSE, נכתבה תוכנה שביצעה את האלגוריתם של CSE. התוכנה נוסתה עבור 10 מונחים. עבור כל מונח חיפוש נאספו 100 תוצאות מגוגל ובסך הכל 1000 תוצאות חיפוש. כאשר עבור כל תוצאה נאסף מידע מיאהו אודות הקטגוריה של התוצאה. ניתוח של התוצאות מלמד על כך שלמרות הפער הרב בסדרי הגודל בין גוגל למדריך יאהו ניתן לקבל מידע מיאהו על מרבית התוצאות מגוגל. בניסוי שנערך, כרבע מ 1000 תוצאות בגוגל

הופיעו ביאהו במדויק. נתון זה עלה ל 40% כאשר נבדקו רק את 100 התוצאות הראשונות שנחשבות איכותיות יותר. כאשר נמנו גם התאמות החלקיות שהשוו את כתובת המתחם (domain) של התוצאות נמצא מידע עבור 70% מהתוצאות. ניתוח כתובת ה URL של התוצאות, מצביע על כך שלמעלה משליש מהתוצאות שנתקבלו מגוגל הן שם המתחם ו 7% מהתוצאות שנמצאו מיאהו הן עמוד פנימי באתר. ממצא זה, מחליש את הטענה הגורסת, שבגלל שמנוע החיפוש מאחזר עמודים, לעומת מדריך האתרים שמסווג אתרים לא ניתן ללמוד ממדריך האתרים על מנוע החיפוש. בבדיקה אחרת נבדקו ידנית 20 תוצאות מתוך 1000 שנתקבלו בניסוי, כשהמטרה היא לברר האם המידע שנתקבל מיאהו מסייע לסווג את התוצאות. הממצאים הם שב- 70% מהמקרים, הקטגוריה שנמצאה במדריך, אכן תורמת לסווג נכון של התוצאות ממנוע החיפוש. וב- 55% מהמקרים, לא רק שהקטגוריה מסווגת נכון את התוצאות, אלא, שהיא מסייעת ממש באחזור המידע, בכך שהיא מוסיפה מידע נוסף ששופך אור על התוצאה שנמצאה. ממצאים אלו מלמדים על כך שהמידע ממדריך האתרים יכול לסייע לסווג את תוצאות החיפוש.

בניסוי נוסף שנערך במחקר נבדק המידע הקטגורי שנתקבל מיאהו. ממצאי המחקר הראו שבלמעלה מ 70% מהקטגוריות שנתקבלו אין זהות בשם הקטגוריה, עובדה זו, הקשתה על פעולת הקיבוץ לפי קטגוריות. לכן הוחלט שהמידע הקטגורי לפיו יעשה הקיבוץ הוא לא על סמך שם הקטגוריה אלא על סמך מיקומה היררכי של הקטגוריה. באופן זה, מוצאה שמיקום הקטגוריה שלה בעץ הקטגוריות הוא News\_and\_Media/Newspapers תקובץ יחד עם תוצאה שמיקומה הוא News\_and\_Media/Magazines .

כדי לנצל את יתרונותיו של CSE וכדי לממש את מטרות המחקר, פותח ממשק שדומה במקצת לממשק של מנוע אשכול, תוך ניצול המבנה היררכי של מדריך יאהו והתאמתו למטרותינו. בצורה זו, רוכזו התוצאות תחת קטגוריות קבועות של נושאים. כשכל בחירה של קטגוריה על ידי המשתמש, חושפת בפניו תוצאות הקשורות לקטגוריה שנבחרה ובמקביל נחשפים בפניו תתי קטגוריות מתאימות. באופן זה המשתמש נחשף במהירות רק לתוצאות מבוקרות בנושא שמעניין אותו, ללא צורך לברור אותן מבין כלל התוצאות. כמו כן, המשתמש נחשף לתוצאות שאילולא המערכת, ספק רב אם היה מגיע אליהם, משום שבמנוע חיפוש סטנדרטי בעל רשימת דירוג, הן היו מופיעות במקומות מרוחקים יחסית. היררכיה של יאהו לא הועתקה במלואה, אלא רק באופן חלקי למערכת CSE מאחר והתברר שהיא עמוקה ורחבה למדי.

המערכת פותחה ויושמה במלואה בסביבת פיתוח של Microsoft.Net 2005 בשפת C#.

המערכת רצה בסביבת Windows 2003 וניתן לגשת אליה בכתובת  
<http://67.192.165.98/cse>.

כדי למדוד, להשוות ולהעריך את CSE הוגדרו מדדים שמוודים את הזמן וכמות התוצאות שעל המשתמש היה לעבור עד למציאת המידע הרצוי. לאחר מכן בוצע ניסוי שהשווה את ציוני המדדים של שלושה מנועי חיפוש: גוגל, carrot2, ו CSE שנתקבלו עבור 9 מונחים מתחומים שונים. תוצאות הניסוי מלמדות על כך שעבור המשתמש הממוצע מערכת ה-CSE מספקת תוצאות טובות יותר מאשר גוגל. CSE סיפקה את המידע המבוקש, בממוצע, פי חמש מהר יותר ועל המשתמש היה לבדוק חמישית מהתוצאות בלבד לעומת גוגל. רק כאשר המשתמש רוצה ומסוגל לבנות מונח חיפוש מפורט המורכב משלוש מילים ויש לו ידע בתחום, החיפוש בגוגל עדיף. CSE אף הציג תוצאות טובות יותר ממנוע האשכול carrot2 כאשר למחפש היה חוסר ידע על מונח החיפוש.

למחקר זה, תרומות מחקריות ויישומיות בתחום מנועי החיפוש ובתחום הערכה של מנועי החיפוש. המחקר מציע אב טיפוס למנוע חיפוש חדשני שמשלב את היתרונות של מנוע חיפוש ומדריך אתרים ובכך מבטיח למשתמש תוצאות איכותיות במהירות. לבסוף, המחקר מציע כיצד לשפר את CSE ולהתגבר על החסרונות שנתגלו בו ובכך להפכו למנוע חיפוש מוביל וחדשני בתחום.

מספר מיון בספרייה:

025.04 גוט.מנ תשס"ח

מספר מערכת:

001153582