אחזור מידע

תרגום חופשי של הספר Information Retrieval

של C. J. van RIJSBERGEN

פרק 8 - העתיד

תרגום : ד״ר צבי קופליק

עריכה : עפר דרורי

מחקר עתידי

בפרקים הקודמים ניסיתי לשלב כמה מהכלים היותר מורכבים בהם משתמשים בעיצוב מערכת אחזור מידע ניסיונית. רבים מהכלים עצמם הם רק בשלב הניסיוני ודרוש עוד מחקר, לא רק לפתח הבנה נכונה שלהם, אלא לעצב את השפעתם על מערכות אחזור מידע קיימות ועתידיות.יתכן ואוכל לציין בקצרה נושאים המזמינים מחקר עתידי.

1. סיווג אוטומטי

הוכחה מוצקה שניתן לטפל באוספי מסמכים בסיווג אוטומטי תעודד עבודה חדשה לדרכי בניית אוספים כאלה. ניתן לצפות שסיווג כזה יעורר עניין גם מסחרי ויחד איתו את התמיכה לפיתוח עתידי. לכן חשוב במידת מה ששימוש בסוג המידע שכבר קיים, שמשתמש בתיאור מסמכים באמצעות מילות מפתח מאפשר את איגודם של מסמכים באוספים גדולים באפן יעיל ואפקטיבי. הכוונה שמחקר נוסף נדרש לזהות דרכים להאצת תהליכי איגוד ללא הקרבת מבנה הנתונים. יתכן וניתן לעצב אלגוריתמים הסתברותיים לתהליכי איגוד אשר יחשבו סיווג מסמכים בממוצע בפחות זמן מאשר המקרה הגרוע ביותר. לדוגמא, יתכן ויתאפשר לצמצם את סדר הגודל של זמן החישוב מ $O(n^2)$, ל$O(n\log n)$, אם כי עבור מספר מקרים פתולוגיים עדיין יידרש זמן חישוב בסדר גודל של $O(n^2)$. דרך נוספת לגשת לבעיה זו של האצת תהליכי האיגוד היא לבחון "כמעט סיווג". יתכן ואפשר לחשב מבני סיווג הקרובים למבנים התיאורטיים הנדרשים אולם מהווים קירובים הניתנים לחישוב ביתר קלות מן האידיאל.

שאלה גדולה שטרם זכתה לתשומת לב רבה היא עד כמה מוגבלת אפקטיביות אחזור המידע על ידי סוג תיאור המסמכים בשימוש. השימוש במילות מפתח לייצוג מסמכים השפיע על הדרך בה עוצבו מערכות סיווג אוטומטיות. יתכן שבעתיד מסמכים ייוצגו בתוך המחשב בדרך שונה לחלוטין. האם קיבוץ מסמכים עדיין יעניין? חושבני שכן.

סיווג מסמכים הוא מקרה מיוחד של תהליך כללי יותר אשר מנסה לנצל קשרים בין מסמכים. קורה שמקדמי אי דמיון שימשו לבטא יחסי מרחק "כאילו". כימות היחסים בדרך זו הוכתב בחלקו על ידי השפה בה תוארו המסמכים. יחד עם זאת, אם היה זה המקרה בו יוצגו מסמכים לא על ידי מילות מפתח אלא בדרך אחרת כלשהי, יתכן ובאמצעות שפה מורכבת יותר, אזי יחסים בין מסמכים יכולים להימדד אחרת. כתוצאה מכך, המבנה בו נייצג יחסים בין מסמכים לא יהיה היראכיה פשוטה אלא במקרה מסוים. במלים אחרות, יש לגשת לאיגוד מסמכים כתהליך למציאת מבנה בנתונים אשר יכול לשמש לשיפור יעילות ואפקטיביות האחזור.

טיעון מקביל לטיעון שהועלה בפסקה הקודמת יכול להינתן לתהליך אוטומטי של סיווג מילות מפתח. השיטות לטיפול במילות מפתח, אשר פותחו ומפותחות, יסייעו ליצירה אוטומטית של מחלקות "יחידות תוכן" אשר ינוצלו במהלך אחזור. סיווג מילות מפתח יישאר על כן מקרה מיוחד.
H. A. Simon בספרו The Sciences of the Artificial, הגדיר מבנה מעניין הקשור למערכת סיווג, הנקרא Nearly Decomposable System. מערכת כזו מורכבת מתתי מערכות כאשר התקשורת (אינטראקציה) בין תתי המערכות שונה בסדר גודל מן התקשורת בתוכן. ההשוואה לסיווג פשוטה, אם נראה קבוצות כתתי מערכות.
סימון ראה את ההגדרה של Nearly Decomposable System כדרך לתיאור מערכות דינאמיות. התכונות הרלוונטיות הן: 1. ב Nearly Decomposable System ההתנהגות בטווח קצר של כל אחת מתתי המערכות בלתי תלויה באחרות ; 2. בטווח הארוך, התנהגותו של כל אחד מן המרכיבים תלויה בהתנהגות המשולבת של כל המרכיבים האחרים. כעת, יתכן שזו השוואה מתאימה להסתכלות בהתנהגות הדינאמית (עדכון ושינוי מילון השפה) של סיווג מסמכים או מילות מפתח.
לאמיתו של דבר, מעט מאוד ידוע אודות התנהגות מבנים סיווגים בסביבות דינאמיות.

2. מבנה קבצים

יעילותה של מערכת אחזור מידע תלויה במבנה הקבצים הנבחר ובשימוש בו.

Inverted Files פופולאריים למדי במערכות אחזור מידע. בוודאי שבמערכות המבוססות על מילות מפתח לא משוקללות, כאשר שאילתות מוגדרות כביטויים בוליאניים Inverted File יכול לתת תשובה מהירה מאוד. לרוע המזל, לא ניתן להשיג התאמה יעילה של Inverted File להתאמת תיאורי

מסמכים ושאילתות מורכבים יותר, דוגמת מילות מפתח משוקללות. מחקר של מבני קבצים היכולים לתמוך ביעילות בתיאורי שאילתות ומסמכים מורכבים עדיין נדרש. יתכן והדרך היחידה להשיג זאת היא על ידי התחלה עם זיווג מסמכים וחקירת מבני הקבצים המתאימים לו. בהתאם יתכן וכיוון שיניב תוצאות הוא חקירת היחסים בין איגוד מסמכים ובסיסי נתונים יחסיים המארגנים את נתוניהם כיחסים n ממדיים.

ישנן בעיות רבות נוספות בתחום זה המעניינות מערכות אחזור מידע. לדוגמא, הארגון הפיזי של מבנים היראַרכיים גדולים המתאימים לאחזור מידע הוא נושא מעניין אחד. כיצד ניתן להגדיר הקצאת נפח אחסון מיטבית להיראַרכיה המיועדת לאחסון על גבי התקן כלשהו בעל מהירויות גישה שונות?

3. אסטרטגיות חיפוש

עד עתה נוסו אסטרטגיות חיפוש פשוטות יחסית. הן נעו בין חיפושים סדרתיים לבין אסטרטגיות מבוססות אגדים, כפי שתואר בפרק 5. צמודה לכל אסטרטגיה מבוססת אגדים היא שיטת ייצוג האגדים. שינוי ייצוג האגדים מביא בדרך כלל לשינוי ההחלטות על חוקי ההחלטה והעצירה של אסטרטגיות החיפוש. דרך אחת, שנראה שטרם נוסתה, מערבת מספר ייצוגי אגדים, כאשר כל אחד נוצר על פי עקרונות אחרים.

אסטרטגיות חיפוש הסתברותיות, גם הן נחקרו רק מעט[1], אם כי אסטרטגיות כאלו נוסו בהשפעה מסוימת בתחומי זיהוי תבניות (Pattern Recognition) ואבחנות רפואיות אוטומטיות. כמובן, בתחומים אלו תיאורי הישויות מפורטים יותר מאשר תיאורי מסמכים באחזור מידע ומשמעות הדבר היא כי יישומן בתחום אחזור המידע יתכן ויחייב תיאור מסמכים מפורט יותר.

בפרק 5 הזכרתי כי נראה שאסטרטגיות חיפוש מלמטה למעלה (Bottom-Up) מוצלחות יותר מאסטרטגיות חיפוש מסורתיות מלמעלה למטה (Top-Down). הדבר מוביל אותי לשער כי עץ פרוש של מסמכים יכול להיות מבנה יעיל להנחיית חיפוש אחר מסמכים רלוונטיים. אסטרטגיית חיפוש המבוססת על עץ פרוש עבור מסמכים תוכל לעשות שימוש במידע אודות תלויות בין מילות המפתח (Index Terms) בעץ. נושא מחקר מעניין הוא לבחון האם על ידי אינטראקציה בין שני עצים פרושים ניתן לשפר את אפקטיביות האחזור.

4. סימולציה

שלושת תחומי המחקר אשר נידונו עד עתה יכולים להיחקר תוך שימוש מוצלח בסימולציה. יש לנו כעת מספיק ידע מפורט המאפשר לנו להגדיר מודל סימולציה של אחזור מידע. לדוגמא, ידוע כי לצורת התפלגות מילות המפתח באוסף מסמכים יש השפעה על אפקטיביות האחזור. מה יכולה להיות השפעת שינוי התפלגויות אילו על סיווג המסמכים או מילות המפתח (Keywords)? יתכן כי ניתן יהיה להגיע למבני קבצים יעילים יותר על ידי חקר ביצועיהם של מבני קבצים שונים תוך סימולציה של התפלגויות שונות של מילות מפתח.

בעיה פתוחה עיקרית היא סימולציה של רלוונטיות. למיטב ידיעתי איש לא הצליח לדמות בהצלחה את המאפיינים של מסמך רלוונטי. ברגע שבעיה זו תיפתר, תיפתח הדרך למחקר השערות דוגמת אשכול והתחברות (Cluster and Association) באמצעות סימולציה.

5. הערכה

זה היה התחום המטריד ביותר באחזור מידע. מוסכם כעת כי נדרשת האפשרות לבצע ניתוח עלות-תועלת או יעילות-אפקטיביות של מערכת אחזור מידע.

בביסוס תיאוריה של הערכה עם תיאורית המדידות, האם אפשרי להגדיר מדידה של אפקטיביות בלי להתחיל מ Precision ו Recall אלא בפשטות תוך שימוש באוסף המסמכים הרלוונטיים ואוסף המסמכים שאוחזרו אם כן, האם ביכולתנו להכליל מדידה שכזו לכלול רמה של רלוונטיות? דרך

---

[1] העבודה המתוארת בפרק 6 מקדמת תרופה למצב זה.

חליפית להגדרת מדידה מן הסוג של "E" יכולה להיעשות במונחי Recall ו Fallout. האם יש בכך יתרון כלשהו?

עד עתה מדידה של אפקטיביות הוכחה כ Intractable בשיטות סטטיסטיות. זאת בעיקר מהעדרו של מודל סטטיסטי מתאים, אולם אין זה אומר שלא קיים כזה![2]

יתכן שקיימים "חוקי" אחזור דוגמת החוק הידוע בדבר יחס החליפין (Trade-Off) בין Precision ו Recall אשר מן הראוי לתקפם באופן ניסויי או תיאורטי. הוכח שיחס חליפין (Trade-Off) אכן, נובע מהנחות בסיסיות יותר אודות מודל האחזור. טיעונים דומים נדרשים לבסס את הגבול העליון לאחזור תחת מודלים מסוימים.

6. ניתוח תוכן

קיים צורך למחקר אינטנסיבי יותר בכל הקשור לייצוג תוכן של מסמכים במחשב.

מערכות אחזור מידע, יישומיות וניסיוניות, מבוססות על מילות מפתח. חלקן נעשו מתוחכמות למדי בשימוש שהן עושות במילות מפתח, לדוגמא, הן יכולות להכיל צורה של נרמול וצורה מסוימת של שקלול. חלקן עושות שימוש במידע הסתברותי למדידת חוזק היחסים בין מילות מפתח או בין תיאורי המסמכים המבוססים על מילות מפתח. נראה שהגענו למגבלת היכולת שלנו לעסוק במילות מפתח כאשר הגדרנו ועשינו שימוש במעט יחסים סמנטיים בין מילות מפתח.

הסיבה העיקרית לגישה פשטנית זו לאחזור מסמכים פשוטה מאוד. מרבית ההוכחות הניסיוניות בעשור האחרון הצביעו על עליונות גישה זו על גישות חלופיות שונות. יחד עם זאת, יש מקום לשיפורים משמעותיים. נראה ששורש אפקטיביות אחזור המידע נעוץ בהתאמת ייצוג המסמכים במחשב. אין ספק שהדבר הוכר כבר בתחילת הדרך אולם ניסיונות להתרחק מייצוג המבוסס על מילות מפתח נחלו הצלחה מועטה באותו זמן. למרות זאת, הייתי רוצה לראות מחקר באחזור מידע הבוחן בשנית את השאלה מה צריך להיות מיוצג במחשב.

הזמן בשל לניסיון נוסף לשימוש בשפה טבעית לצורך ייצוג מסמכים במחשב. ישנה כעת סיבה לאופטימיות מכיוון שקיים ידע רב אודות תחביר (syntax) וסמנטיקה של שפה טבעית. יש לנו כעת מקורות לרעיונות מהתקדמות שהושגה בתחומי מחקר אחרים. בבינה מלאכותית, נעשתה עבודה לקראת הבנה של שפה טבעית על ידי מחשב. נוצרות פרוצדורות מכאניות לעיבוד (והבנה) של שפה טבעית. באופן דומה, בבלשנות פסיכולוגית נחקר המנגנון באמצעותו מבין המוח שפה טבעית. יש להודות כי הדרך בה התקדמות זו ניתנת ליישום באחזור מידע אינה ברורה, אולם ברור כי הן רלוונטיות ויש להתייחס אליהן.

מעולם לא הנחנו כי מערכת אחזור צריכה לנסות ול"הבין" את תכנו של מסמך. מרבית מערכות אחזור המידע כרגע מיועדות לביצוע חיפוש בבלי וגרפי. מסמכים מוגדרים כרלוונטיים על סמך תיאור מלאכותי. אינני טוען כי תכנות מחשב להבין מסמכים אמור להיות תהליך פשוט. מה שמוצע הוא שיעשה ניסיון כלשהו דוגמת Naïve model תוך שימוש ביותר מאשר מילות מפתח בלבד מתוך תוכן כל אחד מן המסמכים באוסף. מערכות המענה לשאלות המתוחכמות יותר עושות משהו דומה. יש להן מודל של עולם התוכן שלהן וביכולתן לענות לשאלות אודותיו וכן לצרף אליו עובדות וחוקים חדשים כאשר אלו מתגלים.

גישה כזו תהפוך את ה"משוב" לכלי מרכזי. משוב, כפי שנעשה בו שימוש כעת, מבוסס על ההנחה שהמשתמש יוכל לבסס הערכה לגבי רלוונטיות מסמך על בסיס נתונים דוגמת כותרת, תקציר ו/או רשימת מונחים על פיהם סווג. הגישה פועלת ברמה מסוימת אך אינה מספקת. אם תוכן המסמך היה מובן על ידי המכונה, הרלוונטיות הייתה מתגלה ביתר קלות על ידי המשתמש. במצב זה כאשר המשתמש מאחזר מסמך, ביכולתו לשאול שאלות פשוטות אודותיו ולהגדיר את חשיבותו בביטחון.

---

[2] חושבני שהמודל של Robertson אשר הוזכר בפרק 7 יכול להיחשב כמודל סטטיסטי סביר

פיתוח עתידי
מחקרים רבים באחזור מידע סבלו מן הקושי להשוות תוצאות אחזור. ניסיונות רבים נעשו עם מגוון רחב של אוספי מסמכים ולעיתים נדירות נעשה שימוש באותו אוסף ביותר ממחקר אחד באותו אופן. על כן תמיד נותר החשד שמא תוצאות מחקר תואמות את הנתונים עליהם הוא נערך ובמידה ויבוצע על נתונים אחרים התוצאות יהיו שונות.

הלקח שצריך להילמד הוא שעבור מחקר חדש חשוב שיהיה בסיס נתונים מוכן. עלה בדעתי אוסף מסמכים בשפה טבעית, כנראה תוך שימוש בכל הטקסט של המסמך. האוסף צריך להיבנות תוך התחשבות במגוון רחב של יישומים ולהיות זמין לכל המעוניין.[3]

נראה כי למערכות אחזור מידע יהיה חלק גדל והולך בקהילה. הן יהיו מקוונות ואינטראקטיביות. החמרה הדרושה לשם כך כבר זמינה, אולם השימוש הנרחב בה ייעשה רק לאחר שתהיה זמינה מסחרית.

התפתחות מרכזית עדכנית היא קישורם של מחשבים ובסיסי נתונים לרשתות. ניתן לחזות שלמשתמשים תהיה גישה לרשתות אלו באמצעות מכשירי טלפון וטלוויזיה כציוד פלט. ההשפעה העיקרית של מערכות אחזור מידע תהיה שהן יהיו פשוטות לשימוש, כאשר הכוונה שהשימוש בהן יהיה בשפה טבעית והן יספקו מידע רלוונטי. מערכת VIEWDATA אשר פותחה על ידי הדואר הבריטי היא דוגמא טובה למערכת אשר תצטרך לענות לדרישות אלו.

בהרחבת אוכלוסיית המשתמשים לכלול משתמשים לא מומחים, סביר להניח כי מערכת אחזור מידע תידרש לספק לא רק ציטוט אלא גם הצגה של הטקסט או חלקו ויתכן שאף לענות על שאלות פשוטות אודות המסמכים שאוחזרו. יתכן שאפילו מומחים ידרשו שמערכת אחזור מידע תעשה יותר מאשר לספק ציטוטים (הצבעות למסמכים).

על מנת לממש את כל המדובר, על מערכת אחזור המסמכים להיות מקושרת ומשולבת עם מערכות אחזור נתונים, על מנת לאפשר גישה לעובדות הקשורות למתואר במסמכים. יישום מיידי נראה כאחזור בתחומי הכימיה או הרפואה. נניח כי משתמש אחזר קבוצת מסמכים אודות תרכובת מסוימת ויתכן שניתן נתון ספקטראלי כלשהו. יתכן והמשתמש ירצה להיוועץ במערכת אחזור נתונים אשר תוכל לתת לו נתונים אודות התרכובת הנידונה, או שיתכן וירצה לגשת באופן מקוון למערכת DENDRAL אשר תיתן לו רשימה של תרכובות אפשריות העקביות עם אותו נתון ספקטראלי. לבסוף, יתכן והוא ירצה לבצע חישוב סטטיסטי כלשהו על הנתונים המצויים במסמכים. לשם כך תידרש לו גישה לתכניות סטטיסטיות.

דוגמא נוספת ניתן למצוא בהקשר של הוראה בסיוע מחשב, כאשר כי זה רעיון טוב לתת לסטודנט גישה למערכת אחזור מסמכים אשר תאפשר לו קריאה נוספת בנושא בו הוא מתעניין כעת. הדחף העיקרי לדוגמאות אלו הוא שיש לתת חשיבות בתיכון של מערכות אחזור מידע לאפשרות לשלבן עם מערכות אחרות.

אף על פי שרישות מחשבים בגודל בינוני נמצא בכותרות ובודדים וארגונים מתחברים לרשתות כדרך להשגת גישה למספר מחשבים, לא ברור שזו תהיה תמיד האסטרטגיה הטובה ביותר. לאחרונה חלה מהפכה בשוק מחשבי הביניים. ניתן לרכוש כיום מחשבים בעלי ביצועים בינוניים בהשקעה נמוכה. מאחר וצינורות המידע ינותבו בעתיד הנראה לעין דרך ספריות, מעניין לחשוב אודות הדרך בה חמרה זולה תשפיע על תפקידן העתידי. ספריות אפשרו גישה למשתמשיהן לבסיסי נתונים גדולים המאוחסנים ומנוהלים במקום מרוחק, יתכן אף שבמדינה אחרת. אפשרות אחת הניצבת בפני הספריות היא זאת שזה עתה הוזכרה, כלומר, ביכולתן להתחבר לרשת גדולה. אפשרות חליפית וגמישה יותר עבורן היא להחזיק, באמצעות מיני מחשב, גישה לאוסף עדכני חלקי שפורסם לאחרונה מתוך אוסף המסמכים. הן יוכלו לשנות אוסף זה מדי פעם. המיני מחשב יהיה חלק מרשת אולם למשתמש תינתן הבחירה האם להפעיל את המערכת הכללית או את המערכת המקומית. המערכת

---

[3] מחקר הממליץ על יצירת אוסף ייחוס שכזה הושלם לאחרונה, ראה Spark Jones and van Rijsbergen, "Information retrieval test collections", Journal of Documentation, 32, 59-75 (1976).

המקומית תותאם לצרכים המקומיים, תכונה שתקנה לה יתרון חשוב. נושאים כמו קבצים אישיים, המכילים מודלי משתמש, יוכלו להישמר על מחשב המיני המקומי. בנוסף, קטלוג הספריה המקומית ואינדקס הנושאים יהיו זמינים בצורה מקוונת, דבר שיהיה שימושי מאוד יחד עם מערכת אחזור מסמכים. משתמש יוכל במהירות לבדוק האם מצויים בספריה עותקי מסמכים שאוחזרו כמו גם ספרים רלוונטיים.

התפתחות חמרה נוספת אשר סביר כי תשפיע על התפתחותן של מערכות אחזור מידע היא שיווקם של מיקרו מעבדים. מאחר ומחירם כה נמוך כעת, רבים שוקלים לתכנן מסופים "אינטליגנטיים" למערכות אחזור מידע, כלומר, כאלו המסוגלים לבצע חלק מן העיבוד הנדרש במקום להותיר הכול למחשב המרכזי. היבט אחד השפעה זו שחלק מן העיבודים "היקרים" יותר יוכלו להתבצע על ידי המסוף, בעוד שבעבר נמנע הדבר.

עם התקדמות האוטומציה בא גם תשלום מס שפתיים לתועלת האפשרית לחברה. לרוע המזל, הטכנולוגיה מתפתחת ומתקדמת בטרם נוכל להעריך האם נרצה בה או לא. במקרה של מערכות אחזור מידע, יש עדיין זמן להעריך ולחקור את השפעתן. אם נחשוב כי מערכות אחזור מידע יתרמו תרומה חשובה עלינו להבהיר מה נספק ומדוע יהווה הדבר שיפור לגישות המקובלות לאחזור מידע.