לקראת רשת משתנים למציאת קשרים בין תוצאות של אלפי מאמרים במדעי החברה

תקציר

"קבירת מידע" – קושי להגיע לתוצאות רלוונטיות ולקשר ביניהן – היא תוצאה של ריבוי המאמרים המתפרסמים מדי שנה במדעה החברה. במאמר זה מתוארת ברמה הקונספטואלית מערכת מידע שמציגה למדענים קשרים בין תוצאות של אלפי מאמרים בתחום ידע נבחר במדעי החברה. המערכת מבוססת על טכנולוגיות עיבוד שפה טבעית ולמידת מכונה, ונועדה לדלות נתונים ממאמרים מדעיים ולארגן אותם ברשת של משתנים (המיוצגים כצמתים) והקשרים הסטטיסטיים שנמצאו ביניהם (המיוצגים כקשתות). המערכת מזהה במאמרים שלשות: משתנה א, קשר סטטיסטי (כגון מתאם או מבחן t) ומשתנה ב. כל שלשה מקושרת גם למאפייניה השונים, כגון תנאי הבדיקה והמחקר. יחידת המידול איננה מייצגת מאמרים אלא לוקחת שלשות ואת מאפייניהן ממאמרים שונים וממזגת אותם יחד לרשת אחת. לאחר שעברה על כל המאמרים, יוצרת המערכת רשת המארגנת את כלל התוצאות הסטטיסטיות מהמאמרים שנבדקו. המערכת המוצעת מאפשרת: (א) גישה קלה לייצוג מסודר של הממצאים שפורסמו בספרות המדעית בעזרת חיפוש טקסטואלי ועל ידי ניווט בגרף ויזואלי של הרשת, (ב) יכולת לבדוק השערות תיאורטיות בהסתמך על כמות גדולה של ממצאים, (ג) איתור מקרים שבהם נבדקו קשרים בין אותם המשתנים במאמרים שונים לצורך ביצוע מטא-אנליזה (ד) כריית נתונים לאיתור קשרים בין קבוצות של משתנים שנבדקו במאמרים שונים, משתנים שהצירוף ביניהם עשוי להוביל למשמעות חדשה. השימושים הללו של הרשת נועדו לשפוך אור חדש על תחומי מחקר ספציפיים במדעי החברה, ולהביא חוקרים לתובנות חדשות על ידי התבוננות בריכוז של תוצאות ממספר מאמרים רב בתחום.

מבוא

"קבירת מידע" (information burying) היא תולדה של ריבוי הפרסומים המדעיים (Mons, 2005). למעלה משני מיליון מאמרים מדעיים מתפרסמים בכל שנה, מהם יותר ממאה אלף מהם במדעי החברה[1], ומספרים אלו ממשיכים לגדול. בשנת 2013 התפרסמו 19,763

[1] National Science Board (2018). *Science and engineering indicators 2018*. Arlington, VA: National Science Foundation. Retrieved from https://www.nsf.gov/statistics/2018/nsb20181/report/sections/academic-research-and-development/outputs-of-s-e-research-publications

מאמרים בפסיכולוגיה,[2] מספר שכמעט הוכפל כעבור שלוש שנים בלבד ל- 39,025 מאמרים ב- 2016. קבירת המידע נובעת מכך ש"מספר המאמרים הרלוונטיים עולה על היכולת האנושית לקריאה, להבנה ולסינתזה" (Mons, 2005, p. 142). בתחום הרפואה והביולוגיה פותחו מערכות להתמודדות עם קבירת המידע, ואלו משתמשות בשילוב של עיבוד שפה טבעית (natural language processing) ואונטולוגיות כדי להציג לרופאים ולחוקרים את המידע הטמון במאמרים אלו בצורה מרוכזת ומתומצתת (Fiszman, Demner-Fushman, Kilicoglu, & Rindflesch, 2009; K. Liu, Hogan, & Crowley, 2011; Y. Liu et al., 2012; Vandenbroeck, Goossens, & Clemens, 2007). ככל הידוע לנו, לא פותחו מערכות דומות למדעי החברה.

מאמר זה מתאר ברמה הקונצפטואלית את מודל הנתונים והמתודולוגיה המחקרית לפיתוח מערכת מידע שתעזור למדענים להתמודד עם בעיית קבירת המידע בתחום מדעי החברה. המערכת המוצעת מתבססת על טכנולוגיות עיבוד שפה טבעית ולמידת מכונה (machine learning) עדכניות כדי לארגן את המידע בתחום ספציפי במדעי החברה. המערכת בנויה לדלות את הנתונים ממאמרים מדעיים ולארגן אותם באופן אוטומטי למחצה ברשת (גרף) של משתנים (המיוצגים כצמתים) והקשרים הסטטיסטיים שנמצאו ביניהם (המיוצגים כקשתות). המשתנים יכולים להיות המאפיינים הדמוגרפיים של משתתפי המחקר (כגון, גיל ומגדר), מאפייני אישיות על פי מודל ה Big Five (כגון, מוחצנות ופתיחות מחשבתית) והתנהגויות מידע (כגון, התמכרות לטלפונים חכמים, ושעות שימוש יומיות באינטרנט). קשרים סטטיסטיים יכולים להיות מתאמים בין זוגות של משתנים (למשל, מתאם Pearson בין גיל משתתפי המחקר ונטייתם להתמכר לטלפון החכם שלהם) או הבדלים בין ערכים במשתנה התלוי עבור ערכים (קטגוריות, קבוצות) שונים של המשתנה הבלתי תלוי (לדוגמה, גברים מכורים למשחקי מחשב יותר מנשים, דבר שנבדק בבחינת t למדגמים בלתי תלויים). ממצאים ממאמרים שונים המתייחסים לאותם משתנים ספציפיים, מקושרים זה לזה ברשת.

רשת זו מאפשרת למומחים בתחום מגוון של פעולות:

- גישה קלה לייצוג מאורגן של הממצאים שפורסמו בספרות המדעית בעזרת חיפוש טקסטואלי ועל ידי ניווט ברשת (גרף ויזואלי) המייצג את המשתנים והקשרים ביניהם;
- יכולת לבדוק השערות תיאורטיות בהסתמך על כמות גדולה של ממצאים;
- איתור מקרים שבהם נבדקו קשרים בין אותם המשתנים במאמרים שונים לצורך מטא-אנליזה;

[2] National Science Board (2016). Science and engineering indicators 2016. Arlington, VA: National Science Foundation. Retrieved from https://www.nsf.gov/statistics/2016/nsb20161/uploads/1/nsb20161.pdf

- כריית נתונים לאיתור קשרים בין קבוצות משתנים שנבדקו במאמרים שונים, ואשר לצירוף ביניהם עשויה להיות משמעות תיאורטית.

ארבעת הפעולות הללו מאפשרות לחוקרים מבט על במידע שנצבר בתחום, לא רק במאמר הבודד אלא ברמה הגלובלית המקשרת בין כמות גדולה של מאמרים. נזכיר כאן שהמערכת האוטומטית המתוארת כאן עדיין איננה פעילה, והמאמר מציג תיאור קונצפטואלי ובדיקת היתכנות ידנית של מודל הנתונים שעליו היא תתבסס וכן של המתודולוגיה המחקרית שפיתחנו כדי לבנות מערכת זו בעתיד.

בפרקים הבאים נסקור את הספרות העוסקת במערכות להפקת מידע מכמות גדולה של נתונים מדעיים ואת אופן בניית רשת המשתנים. נסביר את האופנים השונים בהם ניתן להשתמש בה, נדגים את השימוש במערכת בעזרת ממצאים ממחקר מקדים שבו איסוף וארגון המידע נעשה בצורה ידנית, ונסכם את המאמר תוך ציון הקווים לעבודה עתידית.

סקירת הספרות

הצורך ב- meta-science רחב יריעה, ולא רק במחקרי עומק צרים, נכון לתחומים מדעיים רבים, והמושג מתקשר גם למושגים כלליים יותר כ- e-science ו- e-research (Bechhofer et al., 2013). כך, e-science הוא מגוון המקורות והמשאבים שהאינטרנט מציע לתמיכה בעשייה המדעית (Jankowski, 2007), בכללם ניהול של שיתוף פעולה ותקשורת בין חוקרים הרחוקים גיאוגרפית זה מזה, פיתוח ושימוש בכלים מבוססי אינטרנט לאיסוף נתונים, ניתוח נתונים והצגה ויזואלית שלהם, שימור הנתונים ומתן גישה אליהם ופרסום תוצאות והפצתן. בשנים האחרונות הפך השימוש בטכנולוגיות אינטרנט לחלק מרכזי בעבודת צוות מדעי. כמו כן, כחלק מהמאמץ להנגיש מידע מדעי (open science) נבנו מאגרי נתונים פתוחים (open data) המכילים נתונים מניסויים שונים (Pontika, Knoth, Cancellieri, & Pearce, 2015), דוגמת microarrays בתחום הרפואה, מאגר המאפשר למדענים ברחבי העולם גישה חופשית למידע גנטי (Brown & Botstein, 1999). מאגרים אלו מאפשרים שימוש חוזר בנתונים לצורך שחזור תוצאות הניסוי ולצורך ניתוחים כולל השוואה והצלבה של ממצאים ממחקרים שונים ומטא-אנליזה (Feichtinger, McFarlane, & Larcombe, 2012). ארגון ידע יכול לסייע על ידי סידור שיטתי של התוצאות הקשורות בבעיה סבוכה למבנה אחיד, ללא קשר לגבולות בין תחומי מחקר שונים (Casillas & Daradoumis, 2012). ככל הידוע לנו, בתחום מדעי החברה אין עדיין מספיק נתונים פתוחים ונגישים המאפשרים עיבוד ומטא-אנליזה ולכן במחקר הנוכחי אנו מציעים דרך לשלוף ולנתח תוצאות כמותיות אשר פורסמו במאמרים אקדמיים. לצורך זה על המערכת

המוצעת במחקר שלנו לחלץ באופן אוטומטי יחסים בין מושגי מפתח ממאמרים מדעיים, בעזרת כריית טקסט (text mining) ועיבוד שפה טבעית.

סקירת ספרות זו תעסוק בקצרה בתחומים אשר בעזרתם נפתח את המערכת: אונטולוגיות, עיבוד שפה טבעית ולמידת מכונה. לאחר מכן נסקור כיצד תחומים אלו שימשו לפיתוח מערכות להנגשת מידע בתחום המחקר הרפואי.

אונטולוגיות

התחומים e-science ו- e-research מוזכרים לעתים קרובות בהקשר של מדעי החברה וביו-רפואה. כדי לאפשר אגירה וניהול של מטא-דטה של ניסויים ושל ממצאים בתחומים מדעיים שונים הוצעו מאגרי מידע מובנים רבים וכן אונטולוגיות. **אונטולוגיה** במדעי המחשב היא ייצוג של ידע משותף לתחום דעת מסוים, ידע המיועד לשימושם של מומחים בתחום ושל סוכנים אוטומטיים (כלומר תוכנות מחשב) לצורך תקשורת, שיתוף בין מערכות וניתוח נתונים (Gruber, 1995). אונטולוגיה מורכבת מקבוצת היגדים (עובדות על תחום הדעת) הבנויים כיחסים בינאריים או שלשות: נושא – נשוא – מושא; הנושא והמושא הם מושגים והנשוא הוא היחס הסמנטי ביניהם (Noy & McGuinness, 2001). אין מגבלה לגבי סוג היחסים הסמנטיים בין המושגים באונטולוגיות. לדוגמה: "מתמטיקה – *היא* – סוג של מדע", "לוגיקה – *היא חלק מ* – מתמטיקה", "פיזיקה – *קשורה ל* – מתמטיקה", "פיזיקאי – *חוקר* – פיזיקה". זאת בניגוד לטקסונומיה ותזארוס, בהם ניתן להשתמש רק במערכת יחסים קבועה מראש בין מושגים. אונטולוגיה מהווה מודל נתונים ובסיס ידע עם ייצוג פורמאלי של תחום דעת נתון. אונטולוגיה סטנדרטית משמשת לאגירת מידע ממקורות שונים, וכל אחד מהמקורות משתמש במושגים וביחסים המוגדרים באונטולוגיה.

ניתן לייצג אונטולוגיות באמצעות רשת או גרף (Nickel, Murphy, Tresp, & Gabrilovich, 2016), שבהם המושגים והמופעים (הדוגמאות הקונקרטיות) מיוצגים על ידי צמתים, והיחס ביניהם מיוצג על ידי הקשתות שמקשרות ביניהם. אונטולוגיות משתמשות בכללים לוגיים וכך תומכות בהיסק אוטומטי של קשרים והיגדים חדשים בהתבסס על המידע החבוי באונטולוגיה. למשל, בהינתן שאלברט איינשטיין הוא פיזיקאי, ושאלברט איינשטיין הוא יהודי ושאלברט איינשטיין זכה בפרס נובל, ניתן להסיק שפיזיקאי יהודי זכה בפרס נובל. בנוסף, ניתן לבצע שאילתות מורכבות כדי לדלות את הידע האונטולוגי, ולקבל תשובות מדויקות וממוקדות. כדי להשיג זאת, פיתח ארגון W3C גוון סטנדרטים ל קידוד פורמאלי,

כגון, OWL (שפת אונטולוגיה לרשת) (McGuinness & Van Harmelen, 2004) ושפות שאילתה לאחזור מידע מאונטולוגיות (כגון SPARQL[3]).

מספר אונטולוגיות גנריות וכלי מידול נלווים הוצעו למדענים כדי לארגן ולשמר את תוצאות המחקר שלהם ולשתף פעולה ביניהם. מערכות אלו תומכות בהקצאת משאבים ובשימוש חוזר בהם לצורך סימולציות (Polhill, Pignotti, Gotts, Edwards, & Preece, 2007); מאפשרות למדענים מתחום הביו-רפואה לשלב בין ניסויים שונים בזמן אמת (Ciccarese et al., 2008); לצבור ולחלוק בין החוקרים השונים תכנים לתוך "אובייקטים של מחקר" (research objects) הכוללים לא רק את הנתונים בהם השתמשו, שיטת המחקר שיושמה וצורת הניתוח, אלא גם את החוקרים המעורבים במחקר (Bechhofer et al., 2013; De Roure et al., 2010); וניתוח מידע תוך המחשת התוצאות עבור סימולציות במדעי החברה (Polhill et al., 2007). ההיעזרות במערכות אלו מלמדת על הפוטנציאל הרב שיש בשימוש באונטולוגיות לצרכים מדעיים. המערכת המוצעת במחקר זה עושה שימוש באונטולוגיות לייצוג משתני מחקר ולייצוג הקשרים הסטטיסטיים ביניהם, כשאחד התחומים שבהם היא נעזרת הוא עיבוד שפה טבעית.

עיבוד שפה טבעית

עיבוד שפה טבעית וניתוח טקסט הם תת-תחומים של מדעי המחשב, והם נועדו לנצל קורפוסים טקסטואליים גדולים במטרה לחלץ ולאחזר מהם, בצורה אוטומטית, מידע וידע מטקסט חופשי שאינו מובנה (Manning & Schütze, 1999). שיטות של עיבוד שפה טבעית כוללות אלגוריתמים לניתוח מורפולוגי וחילוץ צורות בסיס של המילה, ניתוח תחבירי (סינטקטי) של מבנה המשפט ותיוג המלים במשפט לפי חלקי דיבר, ניתוח סמנטי ברמה הלקסיקלית וברמת המשפט. בין השאר, עיבוד שפה טבעית מתמודד עם בעיות של זיהוי מלים וקטעי טקסט דומים או גוררים במשמעות (lexical and textual entailment), פענוח רב משמעות (word sense disambiguation), זיהוי הפניה לאותו אובייקט בקטעי טקסט שונים (co-reference resolution), זיהוי שמות פרטיים (named entity recognition), סיכום ותרגום אוטומטי. להלן נפרט על השיטות לניתוח סמנטי שהן הרלוונטיות ביותר לצורך בניית רשת של משתנים סטטיסטיים.

שתי הגישות המרכזיות העומדות בבסיס של אלגוריתמים רבים לניתוח סמנטי של טקסט הן היסק מבוסס מרחב וקטורים מבוזרים (distributional vector space) (Henderson & Popa, 2016; Kotlerman, Dagan, Szpektor, & Zhitomirsky-

[3] Retrieved from http://www.w3.org/TR/rdf-sparql-query/

Geffet, 2010) וזיהוי תבניות לקסיקליות סינטקטיות או מבנים של פרדיקטים וארגומנטים (Hearst, 1992; Stern & Dagan, 2014).

לפי הגישה הראשונה, היסק מבוסס מרחב וקטורים מבוזרים, משמעות המילה משתקפת מתוך ההקשרים שבהם היא מופיעה בטקסטים. לפיכך, למאגר טקסטים גדול יש לבנות עבור כל מילה אוסף או וקטור של תכונות (לרב, מלים וביטויים המופיעים בקרבת המילה הנתונה בטקסטים, ולפי כך מייצגות את ההקשרים שלה) המאפיינות אותה. כל תכונה מקבלת גם משקל מספרי על סמך פרמטרים שונים כגון מידת שכיחותה בהקשרים הקרובים של המלה הנתונה בטקסט. וקטורים אלו מרכיבים את המרחב המבוזר בו כל וקטור מייצג מילה מסוימת, והנחת העבודה היא שמילים דומות במשמעות ייוצגו על ידי וקטורים דומים. כלומר על סמך חישוב הדמיון בין הווקטורים ניתן לזהות מילים בעלות משמעות דומה (Henderson & Popa, 2016; Kotlerman et al., 2010; ז'יטומירסקי-גפת, 2009).

לפי הגישה השנייה, זיהוי תבניות לקסיקליות, האלגוריתם מחפש הופעות משותפות של זוג מילים בתוך הטקסט בתבניות (דפוסים) לקסיקליות סינטקטיות, וזאת כדי לזהות דמיון סמנטי או קשרים סמנטיים ספציפיים בין זוגות מלים (כגון, מלים נרדפות, כלל/פרט, חלק/שלם או גרירה לוגית). לדוגמה, תבנית המייצגת קשר סיבתי [שם המחלה *נגרמת על ידי* שם החיידק] או תבנית מסוג כלל/פרט [מילה א *היא סוג של* מילה ב]. תבניות אלו נבנות באופן ידני או אוטומטי מתוך קורפוס טקסטואלי גדול (Hearst, 2006; Kozareva & Hovy, 2010; Mirkin, Dagan, & Geffet, 2006; Panchenko et al., 2016; Stern & Dagan, 2014).

לאחרונה, וויטיס ושות' (Wities et al., 2017) הציעו שיטה חדשה לניתוח סמנטי של טקסט, שיטה המשלבת את שתי הגישות שתוארו לעיל. האלגוריתם המוצע בונה גרף (רשת סמנטית) של מילים (הנקראות גם ארגומנטים) וקשרי משמעות (פרדיקטים) המקשרים ביניהן מתוך אוסף מסמכים טקסטואליים. בשלב הראשון מתבצע זיהוי תבניות של ארגומנטים-פרדיקטים בטקסט כדי לזהות קשרי משמעות בין מילים. בשלב השני משתמשים בשיטות של מרחב וקטורים מבוזר כדי לאחד מלים וקשרי משמעות שיש ביניהם דמיון סמנטי והם מנוסחים בצורה שונה במסמכים השונים. מזיהוי זה נוצר ייצוג סמנטי תמציתי לטקסט, ללא חזרות מיותרות על אותן העובדות. המערכת המוצעת משתמשת בשיטה משולבת זו. בנוסף, כפי שנראה להלן, על המערכת גם להפעיל שיטות למידת מכונה.

למידת מכונה

כדי לבנות את המודל הרשתי המוצע מתוך הטקסט של המאמרים המדעיים, יש להשתמש גם בשיטות של למידת מכונה מבוקרת (supervised machine learning). שיטות אלו מסייעות הן בבניית מרחב הווקטורים של תכונות עבור המלים השונות (לדוגמה, משתנים סטטיסטיים) והן לזיהוי תבניות לקסיקליות סינטקטיות המייצגות קשרים בין משתנים במאמרים. למידת מכונה מבוקרת היא גישה ללמידת מכונה בה אוסף הנתונים של המחקר מחולק לאוסף אימון (training dataset) ואוסף בוחן (test dataset). בדרך כלל 20% מהנתונים משויכות לאוסף האימון והשאר לאוסף הבוחן. הדוגמות (יחידות מידע, למשל, מילים, צירופי מילים, מקטעים או משפטים) באוסף האימון עוברות תיוג ידני, והדוגמות באוסף הבוחן הן דוגמות חדשות שאינן מתויגות. האלגוריתם מקבל אוסף דוגמות מתויגות לאימון, כדי ללמוד את המודל המאפיין אותן והאמור לייצג את הנתונים באוסף הנתונים של המחקר כולו (Mohri, Rostamizadeh, & Talwalkar, 2012). לאחר מכן, מופעל המודל שנלמד על מדגם בוחן המכיל דוגמות חדשות לא מתויגות במטרה לנבא את תיוגן באופן אוטומטי. כך, אם יש קטעי טקסט המתייחסים לחוות דעת המשתמשים על מוצרים מסוימים, המטרה של האלגוריתם היא ללמוד את סוג הסנטימנט המובע בהם (חיובי, שלילי או נייטרלי) או לסווג אותם לקטגוריות לפי נושאים נתונים. כדי לבדוק את דיוק התוצאות של אלגוריתם הלמידה משתמשים בשיטת הנפוצה של n-fold cross validation. בשיטה זו, כדי ללמוד, בוחרים כל פעם דוגמות שונות כאוסף אימון ומפעילים את האלגוריתם עליהן. לאחר מכן מפעילים את האלגוריתם על שאר הדוגמות כדי לבחון את דיוק הפעולה שלו, וחוזרים על כך n פעמים עבור חלוקה שונה של הדוגמות לאוסף אמון ואוסף בוחן. למשל, אם n הוא 5, וגודל האוסף האימון הוא 20% בהתאם, אז האלגוריתם יורץ 5 פעמים כל פעם על 20% אחרים מתוך הדוגמות כאוסף אימון, וייבחן על 80% שונות של הדוגמות כאוסף בוחן. הדיוק הממוצע מ5 הרצות האלגוריתם על אוספי בוחן שונים הוא שיקבע את דיוק האלגוריתם.

אלגוריתמים של למידה מבוקרת כוללים מודלים של למידה עמוקה (deep learning) המבוססים על רשתות עצביות בעלות יותר משכבת אמצע אחת (Mikolov, Chen, Corrado, & Dean, 2013). היתרון של טכניקות למידה עמוקות הוא ביכולתן לזהות באופן אוטומטי את התכונות המאפיינות (ההקשרים במקרה שלנו) של מילים, ולבנות מרחב וקטורי מבוזר באופן אוטומטי לחלוטין מתוך הטקסט. מודלים אלו הם בדרך כלל רשתות עצביות דו-שכבתיות שאומנו לשחזר הקשרים לשוניים של מילים. לדוגמה, מודל ה-Word2vec הפופולארי כיום בספרות המדעית לוקח את הקלט שלו – שהוא קורפוס גדול של טקסט – ומייצר מרחב וקטורי כך שלכל מילה ייחודית בקורפוס מוקצה וקטור במרחב. המילים בווקטור ממוקמות במרחב הווקטורי כך שמילים בעלות הקשרים משותפים בקורפוס

ממוקמות קרוב זו לזו במרחב. לאחרונה הוצעה הרחבה של Word2vec לבניית וקטור ביטויים ופסקאות מכל המסמכים (ולא מילים בודדות) המתייחסת ל-paragraph2vec (& Le Mikolov, 2014). מיקולוב ושות' (2013) מצאו שאפשר לזהות באופן אוטומטי תבניות סמנטיות ותחביריות באמצעות חישוב וקטורי. חישוב כזה נדרש במחקר שלנו למשל, לזיהוי תבניות המאפיינות זוגות משתנים וקשרים סטטיסטיים ביניהם במאמרים מדעיים.

נחתום את סקירת הספרות בפרק המשלב את הפרקים הקודמים ומדגים כיצד אונטולוגיות, למידת מכונה ועיבוד שפה טבעית מסייעים להנגשת מאמרים בתחום הרפואה. הנגשה זו מסייעת לרופאים לתת למטופלים את הטיפול המיטבי המתבסס על סיכום של הספרות המדעית העדכנית.

הנגשת מאמרים בתחום הרפואה בעזרת אונטולוגיות, למידת מכונה ועיבוד שפה טבעית

שימוש משולב של אונטולוגיות ועיבוד שפה טבעית קיים כבר בתחום הרפואה. אחת הדרכים המקובלות ביותר להציג נתונים בתחום זה היא מפת סיבתיות שהיא סוג מיוחד של מפת מושגים (concept map). לדוגמה, מפת הסיבתיות שיצרו ונדנברוק ושות' (Vandenbroeck et al., 2007) ממפה מעל 200 משתנים הגורמים להשמנת ייתר בקרב ילדים. דוגמה אחרת היא ניסיון להקל את התהליך הידני של בניית אונטולוגיה על ידי מומחים ולהגביר את האובייקטיביות בבחירת המושגים המתאימים להיכלל באונטולוגיה (,Tsoi Patel, Zhao, & Zheng, 2009). לשם כך, הם פיתחו שיטה אוטומטית לזיהוי מושגים מתאימים להכלה באונטולוגיה על סמך מידת שכיחותם ("ייצוגיותם") בתקצירים של PUBMED. פיצמן ושות' (Fiszman et al., 2009) התייחסו למערכת Semantic MEDLINE אשר נועדה לאפשר לרופאים להתמודד עם כמות המידע העצומה של מאמרים רפואיים ב- MEDLINE, באמצעות סיכום גרפי שהוכן בצורה אוטומטית. הם העריכו את יכולתה של המערכת לזהות התערבות תרופתית מוצלחת ב 53 מחלות, והראו שהשימוש במערכת הגביר בממוצע את הדיוק והאפקטיביות של הטיפולים בהן.

לסיכום, מערכות שונות השתמשו באונטולוגיות כדי לסדר, לשתף ולהציג את ממצאיהן. מערכות אחרות (בעיקר בתחום הרפואה) השתמשו בעיבוד שפה טבעית ובלמידת מכונה כדי לחלץ באופן אוטומטי יחסים בין מושגי מפתח ממאמרים מדעיים. מערכות אלו מציגות מידע ניסויי הלקוח ממאמרים עדכניים, מידע שיש בכוחו לסייע לרופאים לתת טיפול טוב יותר. אולם ככל הידוע לנו, מערכות דומות לאלו אינן קיימות במדעי החברה. מטרת מאמר זה היא לתאר מתודולוגיה ומודל נתונים לבניית מערכת אשר שולפת, מנתחת, ומקשרת לרשת משתנים אחת תוצאות מאמרים ממדעי החברה. בנוסף, המערכת מאפשרת לחוקרים במדעי החברה לחפש ולנווט ברשת זו ולחשוף בעזרת כריית מידע קשרים בין מספר משתנים המתייחסים למספר רב של מאמרים. בניגוד למערכות לניתוח מאמרים רפואיים שסקרנו,

מערכת זו אינה מתמקדת ביחסים הסמנטיים, ביבליומטריים או אסוציאטיביים כלליים בין המשתנים השונים (כגון הקשר בין מחלות ודרכי הטיפול בהן), אלא בקשרים סטטיסטיים בין משתנים שונים. נעבור עתה לתיאור המתודולוגיה ליצירת רשת המשתנים במערכת זו.

מתודולוגיה ליצירת רשת המשתנים

פרק זה מסביר כיצד בונים רשת משתנים ויחסים סטטיסטיים שתייצג תוצאות בתחום מסוים במדעי החברה. המתודולוגיה ליצירת רשת המשתנים שאנו מציעים מורכבת משני חלקים: יחידת עיבוד טקסט אשר תחלץ מידע על המשתנים והקשרים הסטטיסטיים ביניהם, ויחידת מידול הנתונים אשר תבנה מהמידע המחולץ את רשת המשתנים ותציג אותה בגרף.

מבנה הרשת

בהתאם לספרות העוסקת בסטנדרטיזציה של תוצאות מאמרים שונים כהכנה למטא-אנליזה (Borenstein, Hedges, Higgins, & Rothstein, 2009; Higgins & Green, 2005) ברשת המוצעת יש שני סוגים של קשרים סטטיסטיים: קשרי R וקשרי D. **קשר R** הוא קשר של מתאם (קורלציה). זהו קשר דו-כיווני שכן אין לגזור את כיוון ההשפעה מהתוצאה הסטטיסטית. המתאם יכול להיות חיובי (יחס ישר בין המשתנים) או שלילי (יחס הפוך בין המשתנים). לדוגמה, כשמערכת מוצאת משפט, כגון, "נמצא מתאם שלילי בין גיל משתתפי המחקר ורמת ההתמכרות שלהם לטלפונים חכמים ($r = -0.6, p < 0.01$)", המשפט יפורק לשלשה: גיל – *במתאם עם* – התמכרות לטלפונים חכמים, ומאפייני הקשר הם: מתאם שלילי מובהק, ערך מתאם 0.6. במקרים של קשרים מורכבים יותר, כגון רגרסיה, נפרק את הקשרים לשלשות של מתאמים לצורך המידול.

קשר D מראה את השפעתם של ערכים שונים של המשתנה הבלתי תלוי (לדוגמה, גברים ונשים הם ערכים שונים של המשתנה "מגדר") על משתנה תלוי (לדוגמה, "רמת ההתמכרות למשחקי מחשב"). אנו נמצא במאמרים קשרי t, ניתוח שונות חד-גורמי (ANOVA) לשני משתנים, וקשרים שנמצאו בבדיקות post hoc לניתוח שונות חד-גורמי בין מספר משתנים. הקשר בין המשתנים הללו הוא חד-כיווני (למשל, המשפט "רמת ההתמכרות למשחקי מחשב של גברים הייתה גבוהה באופן מובהק מזו של נשים" מראה על השפעה של ערכי המשתנה הבלתי תלוי על ערכי המשתנה התלוי). אנו נהפוך את הממצאים הללו לגודל האפקט (effect size), שבדומה למתאם ערכיו נעים בין מינוס אחד לאחד, כדי ליצור סטנדרטיזציה של עוצמות היחסים ברשת (שכן ערך המתאם משמש אף הוא כגודל אפקט). לגבי שני סוגי הקשרים הללו המערכת תשמור בנוסף על גודל האפקט גם נתונים חשובים אחרים, כגון, אם הממצא הגיע למובהקות וגודל המדגם.

מלבד המשתנים הנמדדים ישירות בתוצאה הסטטיסטית, רשת המשתנים מכילה גם משתנים המתעדים את תנאי הבדיקה ואת תנאי המחקר כולו. למשל, כאשר המשתנה הנבדק הוא מהירות האחזור, תנאי הבדיקה יכולים להיות סוג המידע המאוחזר (למשל, קובץ או הודעת דוא"ל) המכשיר עליו נעשה האחזור (כגון, מחשב או טלפון חכם). תנאי המחקר יכולים להיות אופן הבדיקה (למשל, מניפולציה ניסויית, שאלון או מחקר שדה), הארץ בה נערך המחקר ועוד.

יחידת עיבוד הטקסט

על מנת ליישם את המודל, יחידת עיבוד הטקסט נדרשת לניתוח טקסט בשלוש רמות:

1. בחירת המאמרים הרלוונטיים שניתן לבצע באמצעות חיפושם במאגרים אקדמיים רלוונטיים. לצורך הבחירה יש להסתייע במעגל רחב של מומחים בתחום הספציפי של מדעי החברה אשר יצביעו על כתבי העת המפרסמים מאמרים בתחום. בנוסף, יש לבצע חיפוש במאגרי מידע בעזרת מילות מפתח מהתחום.
2. תהליך של פענוח רב-משמעות כדי לפתור את בעיית הפוליסמיה (משמעויות שונות של אותה מילה בשפה) וכן זיהוי מיטונומיה (שימוש בשמות שונים עבור אותה המשמעות). פענוח זה נדרש כדי לזהות שבמאמרים שונים מדובר במשתנים שונים בעלי שם זהה או באותו המשתנה המופיע בחילוף שם.
3. בניית המודל עצמו.

יחידת עיבוד הטקסט מזהה את התוצאות במאמר ומחלצת את המשתנים והקשרים הסטטיסטיים ביניהם בעזרת אלגוריתמים של עיבוד שפה טבעית ולמידת מכונה בגישה מבוססת תבניות לקסיקליות סינטקטיות שתוארה בסקירת הספרות. המטרה היא לזהות שלשות: משתנה א – *קשר סטטיסטי* – משתנה ב. יחידת הטקסט מתחילה בחיפוש מילות מפתח הקשורות לממצאים סטטיסטיים (כגון "correlation" ו "t test") ולומדת תבניות לקסיקליות סינטקטיות הנמצאות בקרבה למילים אלו מתוך הטקסטים של המאמרים השונים. לאחר מכן, הטקסט והתבניות שנלמדו ישמשו לזיהוי המשתנים והיחסים הסטטיסטיים ביניהם. למשל, יחידת הטקסט לומדת מהדפוסים לזהות את המשתנים שהקשר הסטטיסטי ביניהם הוא מתאם או תלות. לשם כך, יש לעשות שימוש בלמידת מכונה מבוקרת (supervised machine learning) ולפי הצורך גם במודלים של למידה עמוקה שתוארו בסקירת הספרות.

במקרה שלנו, הדוגמות שניתנות כקלט לאלגוריתם למידה הם קטעי טקסט או דפוסים לקסיקליים. מומחה אנושי תייג את הקשר בין המשתנים המופיעים בתוכם, והאלגוריתם יתאמן עליהם, כדי שבהמשך יוכל לזהות את הקשר בין משתנים בקטעי טקסט וכן דפוסים

חדשים שהקשר ביניהם לא ידוע לאלגוריתם מראש. מומחים בתחום הידע בדקו את האלגוריתם, וכתוצאה מהבדיקה נעשו תיקונים ביחידת עיבוד הטקסט כדי להשיג תוצאות מדויקות יותר בניתוחי המאמרים הבאים. תהליך זה חוזר על עצמו פעמים רבות עד שיחידת עיבוד הטקסט תגיע לרמת ביצוע של מומחה בתחום. ניתן להגיע לדיוק גבוה מזה של מומחה בודד על ידי הצלבה בין החלטות של מומחים שונים בניתוח של אותן תוצאות.

נציין שבניגוד לתחום הרפואה, בתחום מדעי החברה שהוא הנושא של המחקר הנוכחי לא קיימים לקסיקונים ואונטולוגיות סטנדרטיים נגישים לחוקרים, ולכן אחת המשימות של המחקר היא לבנות אונטולוגיה כזו בצורה אוטומטית למחצה על סמך ניתוח סמנטי של המאמרים המדעיים. לשם כך, ניתן להפעיל את השיטות האוטומטיות שהוזכרו לעיל לזיהוי מלים הדומות במשמעות וכן זיהוי יחסי גרירה לקסיקליים (Kotlerman et al., 2010; Shwartz & Dagan, 2016), ולשלב אותן עם בקרת מומחים כדי להגיע לדיוק מרבי.

יחידת מידול והצגת המידע

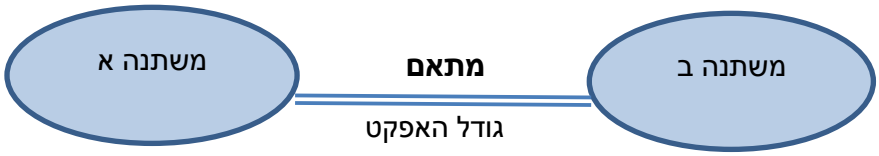
התוצר של יחידת עיבוד הטקסט יכיל אוסף שלשות של משתנה א – *קשר סטטיסטי* – משתנה ב. כל שלשה תקושר גם למאפייניה השונים, כגון, תנאי הבדיקה והמחקר, ולערך התוצאה הסטטיסטית, גודל המדגם והמובהקות כפי שהופיעו במאמר. השלשות לא יקושרו להשערות המחקר התאורטיות שלא יוצגו במודל הרשת. יחידת המידול לא מייצגת מאמרים אלא לוקחת שלשות ומאפייניהן שמקורן ממאמרים שונים וממזגת אותן יחד לרשת אחת. לשם כך, היא מזהה וממזגת לצומת אחד ברשת הופעות שונות של אותו משתנה בתוך אוסף השלשות שתקבל מיחידת עיבוד הטקסט. בין הבעיות בעיבוד השפה שאנו צופים: כינוי אותו משתנה/קשר סטטיסטי בשמות שונים במאמרים שונים, שימוש באותו מונח למשתנים/קשרים שונים, משתנים שאינם מוגדרים, או מוגדרים בדרכים שונות במאמרים שונים, ומידע סטטיסטי המפוזר בחלקים שונים של המאמר או חסר (למשל, הממוצעים וסטיות התקן יכולים להופיע בטבלה נפרדת או לא להופיע כלל, דבר שעלול להפריע לחישוב גודל האפקט). כדי לזהות הופעות של אותו משתנה במקומות שונים באותו מאמר וכן במאמרים שונים יש להפעיל שיטות לזיהוי מלים וקטעי טקסט הדומים או גוררים זה את זה במשמעותם (lexical and textual entailment) (Henderson & Popa, 2016; Kotlerman et al. 2010), שיטות המבוססות על חישוב הדמיון בין הווקטורים של אותם המלים/קטעי טקסט במרחב הווקטורים המבוזרים כפי שתואר בסקירת הספרות. כמו כן, יש להשתמש גם באונטולוגיות קיימות, כגון, WordNet לפענוח רב-משמעות.

משתנה יכול להופיע במאמרים שונים וכן באותו מאמר מספר פעמים בהתייחסות לקשרים הסטטיסטיים שלו עם משתנים שונים. מידת החפיפה בין שלשות יכולה להיות מלאה או חלקית. חפיפה חלקית תקרה כששלשה מכילה קשר בין אותם שני משתנים אולם

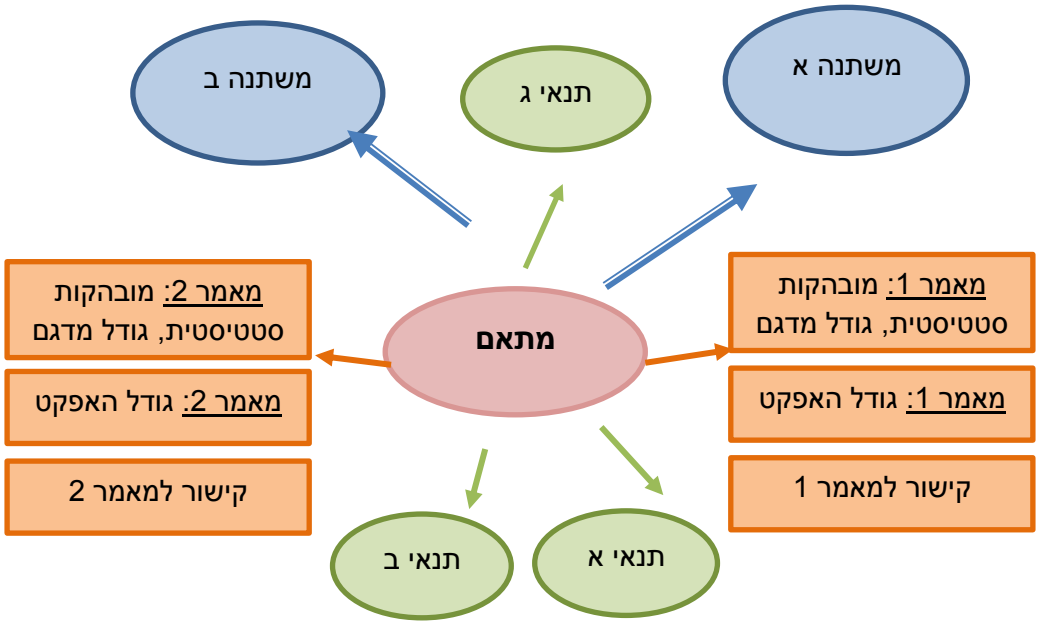
בתנאי בדיקה שונים או קשר סטטיסטי אחר. הדבר מיוצג ברשת כקשתות שונות בין אותם צמתים (משתנים). לדוגמה, מתאם בין מהירות האחזור ובין רמת אוריינות מחשב תיבדק במחקר אחד לגבי קבצים במחשב ובמחקר אחר לגבי דוא"ל בטלפונים חכמים. כאשר גם תנאי הבדיקה יהיו זהים מדובר ביחס של חפיפה מלאה בין השלשות שייוצג כקשת אחת המורכבת ממספר קוים מקבילים ויאפשר לחוקרים לאתר בקלות אזורים אלה לצורך ביצוע מטא-אנליזות.

הרשת מוצגת לחוקרים בשתי רמות של ייצוג נתונים - דו-ממדית ותלת-ממדית (ראו תרשים 1). ברמה הדו-ממדית ניתן לראות את סוג הקשר בין שני המשתנים (R או D) ואת גודל האפקט של התוצאה הסטטיסטית. בתצוגה תלת-ממדית ניתן לראות נתונים נוספים, כגון, חוזק המובהקות, מספר הנבדקים, תנאי הבדיקה והמחקר, התוצאה כפי שהופיעה במאמר וקישור לטקסט המאמר עצמו. התצוגה הדו-ממדית מתאימה יותר להתבוננות במפה הכללית בהתאם לפרדיגמת "הקריאה מרחוק" (distant reading) המכילה משתנים רבים, והתצוגה התלת-ממדית תאפשר "לקרוא מקרוב" ולהתמקד בקשר בין שני משתנים ספציפיים (close reading) (Moretti, 2005). במצב ברירת המחדל המערכת מציגה רק את הקשרים המובהקים בין המשתנים, אולם ניתן לראות גם קשרים המייצגים ממצאים לא מובהקים (למשל, לצורך הכללה ומטא-אנליזה). לאחר שהמערכת תסיים לעבור על כל המאמרים שהוקצו לה בזמן נתון, תיווצר רשת המתייחסת לתחום ידע מסוים והיא תארגן את כלל התוצאות הסטטיסטיות מהמאמרים השונים. נפנה עתה לשימושיה של רשת זו.

ייצוג דו-ממדי



ייצוג תלת-ממדי



תרשים 1: קטע מרשת המשתנים המדגים את מודל יחס המתאם בין זוג משתנים בשתי רמות הייצוג: דו-ממדי ותלת-ממדי. המשתנים מיוצגים על ידי אליפסות כחולות, והקשרים ביניהם מיוצגים על ידי קשתות/חצים כחולים בהתאם. היחס נבחן בשני מאמרים שונים בחפיפה מלאה של תנאים ולכן המשתנים מחוברים בשני קוים מקבילים. ברמה התלת-ממדית הקשר בין המשתנים מיוצג כצומת מיוחד בצבע ורוד אשר מחבר יחד את כל הישויות (משתנים, תנאים ומאפיינים) הקשורות ליחס המתאם הנתון. תנאים מיוצגים כאליפסות ירוקות ומאפייני הקשר מכל אחד מהמאמרים מיוצגים כמלבנים כתומים.

השימוש ברשת המשתנים

מערכת מידע זו, המבוססת על רשת המשתנים, מאפשרת לחוקרים בתחום מגוון רחב של סוגי שימוש. חלק מהן ממוקדות מטרה (top-down) – מנשק משתמש נוח עבור גישה קלה לייצוג של ממצאים מדעיים שפורסמו ובדיקת השערות תאורטיות על סמך כמות גדולה של ממצאים. וחלק מהן מקבלות את המיקוד שלהן מהנתונים עצמם (bottom up) – הצבעה על מקומות מתאימים למטא-אנליזה וכריית מידע מהנתונים שהצטברו ברשת.

גישה קלה לממצאים

המערכת מאפשרת לאתר ולגלות קשרים בין שני משתנים או קבוצה קטנה של משתנים שהופיעו בספרות המדעית בתחום מסוים, זאת בעזרת חיפוש טקסטואלי או על ידי ניווט בגרף ויזואלי של רשת המשתנים. לדוגמה, היא מאפשרת לאתר את כל המאמרים שבדקו קשר בין התמכרות לטלפונים ובין ציונים בבית ספר או לגלות את כל המשתנים הקשורים באופן מובהק להתמכרות לטלפונים חכמים. היכולת לראות רק את הנתונים הרלוונטיים ביותר בצורה מאורגנת תוכל בין השאר לסייע להעריך את מידת החדשנות של הצעת מחקר, בקשה למענק או פרסום העומד לשיפוט. המערכת גם מאפשרת גישה מהירה לתוצאות כפי שהופיעו במאמר המקורי, ולמאמר עצמו למקרה שהמשתמשים ירצו להרחיב ולראות את ההקשר והתוקף של הממצאים.

בדיקת השערות תיאורטיות

תאוריות מייצרות השערות קונקרטיות לגבי ממצאים אמפיריים. הרשת איננה מייצגת תאוריות, ומידול הממצאים באשר לתיאוריות הוא ניטראלי ככל הניתן. אולם, המערכת מאפשרת למדענים לעמת תיאוריות מנוגדות תוך בחינה מקיפה של תוצאות ממאמרים רבים. למשל, בתחום של אינטראקציית אדם-מחשב יש שתי גישות מנוגדות. האחת גורסת שכדאי להפוך כל פעולה שניתן לאוטומטית, וכך להקל על המשתמשים ולמנוע טעויות. על פי הגישה השנייה, כדאי להשאיר את השליטה בידי המשתמשים, משום שתחושת השליטה חשובה להם וכן בגלל שאנשים מתפקדים טוב יותר ממחשבים בתנאים לא צפויים, כגון תנאי חירום. המערכת מאפשרת למדענים בתחום לבדוק במהירות כמות עצומה של מחקרים אמפיריים שבדקו השערות אלו, ואולי אף להכריע על סמך ממצאים אמפיריים מהם התנאים שבהם עדיף לבחור באוטומציה ובאלו בשליטת המשתמש.

מטא-אנליזה

מטא-אנליזה היא שיטה סטטיסטית לסיכום מאמרים שונים שבהם נבדק הקשר בין אותם משתנים (Borenstein et al., 2009). המטא-אנליזה מגדילה את העוצמה הסטטיסטית ואת

דיוק התוצאות ומאפשרת הכללה טובה יותר של ממצאי המחקרים השונים לגבי כלל האוכלוסייה (Higgins & Green, 2005). הנחת היסוד של המטא-אנליזה היא שיש אמת אחת מאחורי הממצאים השונים והשוני ביניהם נובע מטעות סטטיסטית. אשר על כן, התוצאות השונות אמורות להתפלג בהתפלגות נורמלית (עם תיקונים עבור "תוצאות מגירה", כלומר בהנחה שחלק מהתוצאות לא מגיעות למובהקות ולכן אינן מפורסמות), ואם כך, ניתן לחשב ממוצע משוקלל של גודל האפקט של כל התוצאות ולייצג תוצאות מאמרים רבים בעזרת מספר אחד. אבל, אם תוצאות המאמרים השונים אינן מתפלגות בצורה נורמלית ו/או השונות ביניהן גדולה מדי, עורכי המטא-אנליזה מחפשים את הגורמים (moderators) היוצרים **שונות שיטתית** בין התוצאות, ומנסים לפלג בעזרתם את תוצאות המאמרים לקטגוריות שונות שלכל אחת מהן ממוצע משוקלל משלה. בתהליך המטא-אנליזה יש חשיבות רבה לשיקול הדעת האנושי, החל מהקריטריונים לקבלת מאמרים ולפסילתם (למשל, בשל בעיות מתודולוגיות), דרך אופן הטיפול בנתונים חסרים (למשל, ישנה אפשרות לפנות ישירות לחוקרים שכתבו את המאמר ולבקש מהם נתונים החסרים לחישוב גודל האפקט, כגון, סטיות תקן), וכלה במציאת ה- moderators הרלוונטיים הגורמים לשונות השיטתית (Borenstein et al., 2009; Higgins & Green, 2005).

המערכת המוצעת במאמר זה מאפשרת לחוקרים למצוא קשרים בין אותם זוגות משתנים בשלשות חופפות אשר נבדקו במאמרים שונים. אין היא מנסה לבצע מטא-אנליזה באופן אלגוריתמי, אלא רק לסייע לחוקרים בתחום לבצע פעולה זו באופן ידני. המערכת עוזרת לחוקרים למצוא קשרים בין משתנים אשר נבדקו במחקרים רבים, מציגה להם את הנתונים הללו בצורה ויזואלית ומספקת להם קישורים למאמרים עצמם לצורך העמקה והפעלת שיקול דעת הנדרש במטא-אנליזה. לאחר ביצוע המטא-אנליזה יכולים החוקרים בתחום להזין את תוצאותיה בחזרה למערכת כדי לשפר את רשת הקשרים בין המשתנים.

כריית נתונים

המערכת מאפשרת שימוש בשיטות מתקדמות לניתוח נתונים רבים (big data), במטרה לזהות מערכות קשרים סבוכות בין משתנים, לגלות מגמות משתנות במחקר ולהבחין בתת-תחומים אשר טרם הבחינו בהם. כאשר תוצאות המחקרים מאורגנות ברשת מובנית מתאפשרת גישת מחשב (machine access) לנתונים אלו, וגישה זו מאפשרת ניתוחים של כמות גדולה של תוצאות. המערכת מבצעת כריית נתונים ברשת המשתנים המוצגת כגרף ומחשבת מאפיינים סטנדרטיים של הרשת כגון: מידת המרכזיות של המשתנים השונים, מידת הקישוריות שלהם, הצפיפות והדלילות שלהם (Bonacich, 1987; Borgatti, 2005). ניתוח זה מאפשר להבחין ב"קליקות" – קבוצות משתנים בעלי קשרים חזקים ביניהם וכן לגלות את המשתנים המרכזיים (בהם יש שימוש רב ביותר בספרות), משתנים עם הכמות

הרבה ביותר של קשרים וקשרים אשר טרם נבדקו בין משתנים (אזורים דלילים ברשת). כמו כן מאפשרת המערכת , וכן את האפשרות היסק של קשרים חדשים, למשל, על סמך תכונת הטרנזיטיביות, כך שאם יש קשר חזק בין משתנה א למשתנה ב ובין משתנה ב למשתנה ג, ייתכן שיש קשר גם בין משתנה א למשתנה ג. אף כי ייתכן שרבות מתוצאות כריית המידע יהיו טריוויאליות או חסרות משמעות, מומחים בתחום יוכלו לפרש חלק מהתוצאות בדרכים אשר ירחיבו תאוריות קיימות או אפילו יסיעו בפיתוחן של תיאוריות חדשות שאפשר יהיה לבדוק באופן אמפירי.

המתודולוגיה ליצירת רשת משתנים ואופן השימוש בה מודגמים בניסוי בדיקת ההיתכנות שבנספח 1. בנספח מתואר ניסוי פיילוט שבמהלכו ניתחנו באופן ידני חמישה מאמרים בתחום ניהול מידע אישי (personal information management).

סיכום

במאמר זה הצגנו מודל נתונים (data model),מתודולוגיה ובדיקה מקדמית לבניית מערכת המכילה רשת של משתנים וקשרים סטטיסטיים ביניהם, כדי לאפשר לחוקרים בתחומים שונים של מדעי החברה לקבל מידע מאורגן לגבי אלפי מאמרים בתחום אותו הם חוקרים. המערכת שהצענו משתמשת בעיבוד שפה טבעית ובלמידת מכונה מבוקרת, כדי לזהות תוצאות סטטיסטיות ולפרק אותן לשלשות של משתנה א – *קשר סטטיסטי* – משתנה ב. לאחר מכן, השלשות ומאפייניהן מקושרות וממוזגות לרשת הכוללת משתנים וקשרים רבים, אותה ניתן להציג באופן ויזואלי כגרף. רשת זו תאפשר גישה קלה לממצאי המחקרים לצורך בדיקת תיאוריות, ביצוע אנליזות, וכן לצורך כריית נתונים – דבר שאינו אפשרי כיום. בהמשך הדרך אנו מתכוונים ליישם ולבחון בפועל את המודל ואת המתודולוגיה המוצעת עבור בניית מערכת אוטומטית לתחום ספציפי במדעי החברה. ברור לנו שעומדים בפנינו אתגרים גדולים מאוד בבניית המערכת והשימוש בה. אולם גם אם ההצלחה בבניית המערכת תהיה חלקית בלבד תהיה בכך חשיבות גדולה כצעד ראשון לקראת מערכת שתעזור לחוקרים להתמודד עם בעיית קבירת המידע.

תודות

אנו מודים לשיר הלל על עזרתה באיסוף הנתונים למחקר זה.

מקורות

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., . . . Dunlop, I. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems, 29*(2), 599-611.

Bergman, O., Beyth-Marom, R., Nachmias, R., Gradovitch, N., & Whittaker, S. (2008). Improved search engines and navigation preference in personal information management. *ACM Transactions on Information Systems, 26*(4), 1-24. doi: 10.1145/1402256.1402259

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology, 92*(5), 1170-1182.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. New Jersey: John Wiley & Sons, Ltd.

Borgatti, S. P. (2005). Centrality and network flow. *Social networks, 27*(1), 55-71.

Brown, P. O., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature genetics, 21*(1s), 33.

Casillas, L., & Daradoumis, T. (2012). An Ontological Structure for Gathering and Sharing Knowledge among Scientists through Experiment Modeling *Collaborative and Distributed E-Research: Innovations in Technologies, Strategies and Applications* (pp. 165-179): IGI Global.

Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., & Clark, T. (2008). The SWAN biomedical discourse ontology. *Journal of biomedical informatics, 41*(5), 739-751.

De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., . . . Missier, P. (2010). *The evolution of myexperiment.* Paper presented at the e-Science (e-Science), 2010 IEEE Sixth International Conference on.

Feichtinger, J., McFarlane, R. J., & Larcombe, L. D. (2012). CancerMA: a web-based tool for automatic meta-analysis of public cancer microarray data. *Database, 2012*.

Fiszman, M., Demner-Fushman, D., Kilicoglu, H., & Rindflesch, T. C. (2009). Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of biomedical informatics, 42*(5), 801-813.

Fitchett, S., & Cockburn, A. (2015). An empirical characterisation of file retrieval. *International Journal of Human-Computer Studies, 74*, 1-13.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies, 43*(5-6), 907-928.

Hearst, M. A. (1992). *Automatic acquisition of hyponyms from large text corpora.* Paper presented at the Proceedings of the 14th conference on Computational linguistics-Volume 2.

Hearst, M. A. (2006). Clustering versus faceted categories for information exploration. *Communications of the ACM, 49*(4), 59-61. doi: http://doi.acm.org/10.1145/1121949.1121983

Henderson, J., & Popa, D. N. (2016). A vector space for distributional semantics for entailment. *arXiv preprint arXiv:1607.03780*.

Higgins, J. P., & Green, S. (2005). Cochrane handbook for systematic reviews of interventions: version.

Jankowski, N. W. (2007). Exploring e-science: an introduction. *Journal of Computer-Mediated Communication, 12*(2), 549-562.

Kotlerman, L., Dagan, I., Szpektor, I., & Zhitomirsky-Geffet, M. (2010). Directional distributional similarity for lexical inference. *Natural Language Engineering, 16*(4), 359-389.

Kozareva, Z., & Hovy, E. (2010). *A semi-supervised method to learn and construct taxonomies using the web.* Paper presented at the Proceedings of the 2010 conference on empirical methods in natural language processing.

Le, Q., & Mikolov, T. (2014). *Distributed representations of sentences and documents.* Paper presented at the International Conference on Machine Learning.

Liu, K., Hogan, W. R., & Crowley, R. S. (2011). Natural language processing methods and systems for biomedical ontology learning. *Journal of biomedical informatics, 44*(1), 163-179.
Liu, Y., Bill, R., Fiszman, M., Rindflesch, T., Pedersen, T., Melton, G. B., & Pakhomov, S. V. (2012). *Using SemRep to label semantic relations extracted from clinical text.* Paper presented at the AMIA annual symposium proceedings.
Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999): MIT Press.
McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C recommendation, 10*(10), 2004.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* Paper presented at the Advances in neural information processing systems.
Mirkin, S., Dagan, I., & Geffet, M. (2006). *Integrating pattern-based and distributional similarity methods for lexical entailment acquisition.* Paper presented at the Proceedings of the COLING/ACL on Main conference poster sessions.
Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*: MIT press.
Mons, B. (2005). Which gene did you mean? *BMC bioinformatics, 6*(1), 142.
Moretti, F. (2005). *Graphs, maps, trees: abstract models for a literary history*: Verso.
Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE, 104*(1), 11-33.
Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology: Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, Stanford, CA.
Panchenko, A., Faralli, S., Ruppert, E., Remus, S., Naets, H., Fairon, C., . . . Biemann, C. (2016). *TAXI at SemEval-2016 Task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling.* Paper presented at the Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).
Polhill, J. G., Pignotti, E., Gotts, N. M., Edwards, P., & Preece, A. (2007). A semantic grid service for experimentation with an agent-based model of land-use change. *Journal of Artificial Societies and Social Simulation, 10*(2), 2.
Pontika, N., Knoth, P., Cancellieri, M., & Pearce, S. (2015). *Fostering open science to research using a taxonomy and an eLearning portal.* Paper presented at the Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business.
Shwartz, V., & Dagan, I. (2016). Path-based vs. distributional information in recognizing lexical semantic relations. *arXiv preprint arXiv:1608.05014*.
Stern, A., & Dagan, I. (2014). *Recognizing implied predicate-argument relationships in textual inference.* Paper presented at the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).
Tsoi, L. C., Patel, R., Zhao, W., & Zheng, W. J. (2009). Text-mining approach to evaluate terms for ontology development. *Journal of biomedical informatics, 42*(5), 824-830.
Vandenbroeck, P., Goossens, J., & Clemens, M. (2007). *Tackling Obesities: Future Choices — Obesity System Atlas*. London, UK: Retrieved from https://www.gov.uk/government/publications/reducing-obesity-obesity-system-map.
Wities, R., Shwartz, V., Stanovsky, G., Adler, M., Shapira, O., Upadhyay, S., . . . Dagan, I. (2017). *A consolidated open knowledge representation for multiple texts.* Paper presented at the Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics.
ז'יטומירסקי-גפת, מ. (2009). מהן מילים "דומות במשמעותן" וכיצד הן מסייעות לאחזור מידע. מידעת, 5, 59-69.

נספח 1 – ניסוי בדיקת היתכנות של בניית רשת משתנים וקשרים סטטיסטיים

על מנת לבדוק היתכנות של המודל ושל הגישה המוצעת לבניית רשת של משתנים סטטיסטיים ביצענו ניסוי פיילוט. במהלך הפיילוט ניתחנו באופן ידני חמישה מאמרים בתחום ניהול מידע אישי (personal information management). ניהול מידע אישי, הוא תחום מחקר העוסק בהתנהגות מידע בה המשתמש מסדר פריטי מידע (כגון קבצים, דוא"ל ומועדפים) כדי לאחזר אותם בעצמו בזמן מאוחר יותר. מידע על חמשת מאמרים אלו מופיע בטבלה 1. כתוצאה מהניתוח חולצו 20 משתנים, 27 ערכים שונים של משתנים אלו ו- 20 מבחנים סטטיסטיים לגבי קשרים ביניהם (מהם 17 קשרי D ו-7 קשרי R). לשם הדגמה נתמקד במאמרים 4 ו- 5. במאמרים 4 ו-5 זוהו שלושה משתנים משותפים: זמן אחזור (retrieval time), שיטת אחזור (retrieval method) ותוצאת אחזור (retrieval outcome). למשתנים אלו היו גם ערכים ספציפיים, כגון navigation, search, failure, success. משתנים אלו וערכיהם הופיעו במאמרים בשמות שונים (פירוט בהמשך).

טבלה 1 – פרטי המאמרים בפיילוט.

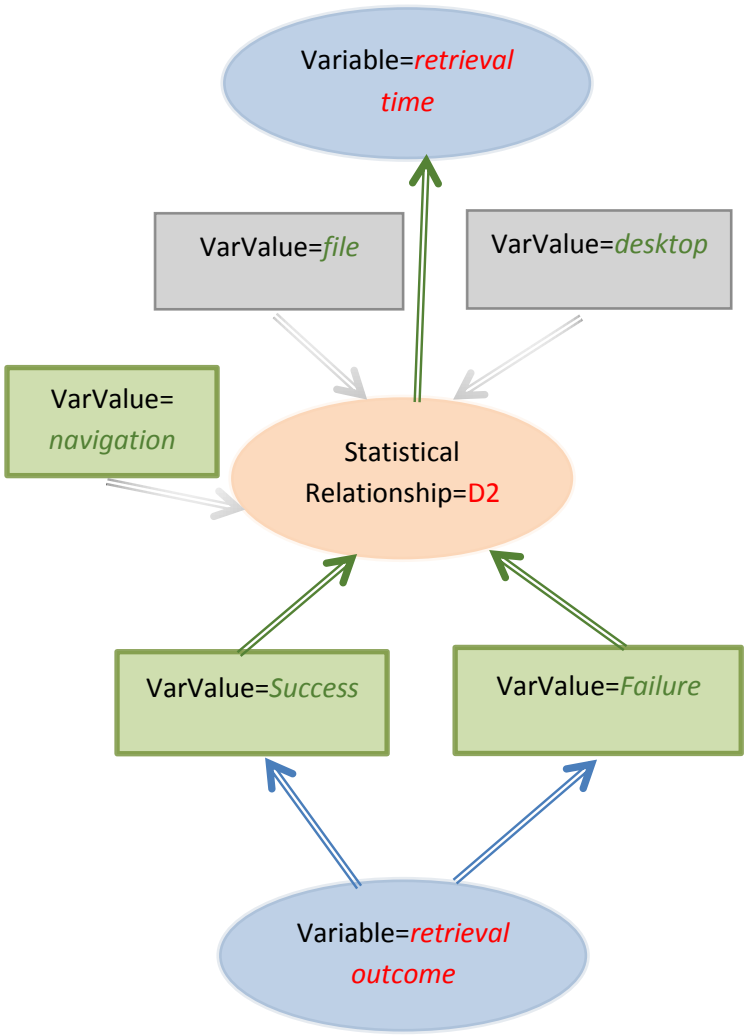
Paper_id	Paper_name	Authors	Year	Journal
1	PIM and personality: what do our personal file systems say about us?	Charlotte Massey, Sean Ten Brook, Chaconne Tatum, Steve Whittaker	2014	Proceedings of the SIGCHI Conference on Human Factors in Computing Systems
2	Folder Versus Tag Preference in Personal Information Management	Ofer Bergman, Noa Gradovitch, Judit Bar-Ilan, Ruth Beyth-Marom	2013	Journal of the Association for Information Science and Technology
3	An Investigation of Memory for Daily Computing Events	Mary Czerwinski, Eric Horvitz	2002	People and computers XVI-memorable yet invisible
4	The use of attention resources in navigation versus search	Ofer Bergman, Maskit Tene-Rubinstein, Jonathan Shalom	2013	Personal and ubiquitous computing
5	The Effect of Folder Structure on Personal File Navigation	Ofer Bergman, Steve Whittaker, Mark Sanderson, Rafi Nachmias, Anand Ramamoorthy	2010	Journal of the Association for Information Science and Technology

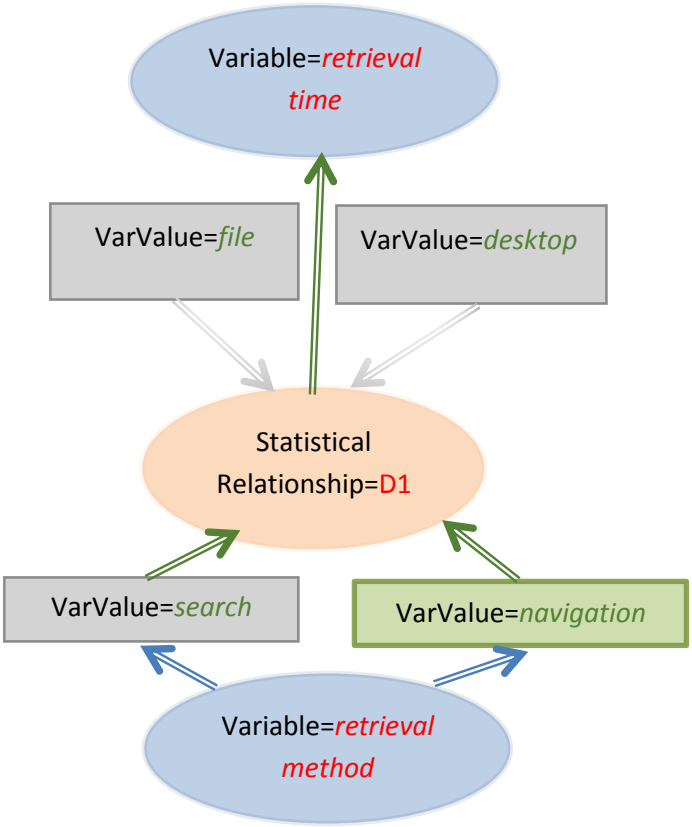
מאמר 4 בחן באמצעות ניסוי משתמשים, הבדלים בין השימוש בשיטות האחזור ניווט וחיפוש לצורך הגעה לקבצים במחשב ובפרט בחן את השפעת שיטת האחזור על זמן

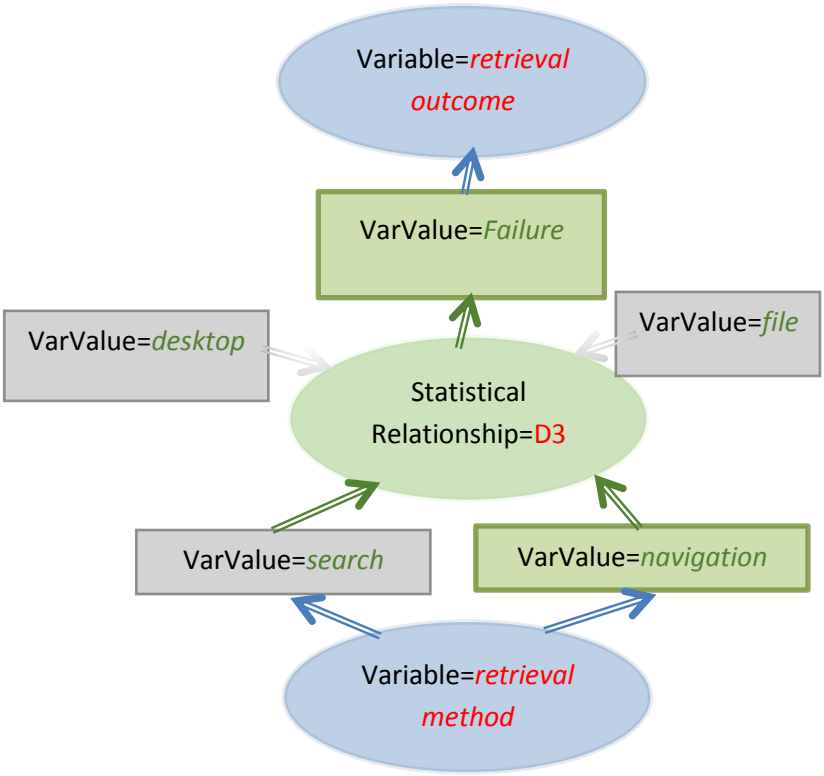
האחזור ותוצאת (מידת הצלחת) האחזור. מחשב (סוג מכשיר) וקובץ (פורמט המידע המאוחזר) אלו ערכי משתנים שלא נבחנו במאמרים, אך הם מהווים תנאים לקשרים שנבחנו במחקר זה (החלק הימני של תרשים 2).

במאמר 5 המטרה הייתה לבחון את השפעת תוצאות האחזור השונות (כגון, הצלחה וכישלון) על זמן האחזור של קבצים בשיטת הניווט במערכת הקבצים במחשב הנייד. למשל, נמצא בניסוי שנערך במחקר המתואר, כי הצלחה ישירה (direct success) של אחזור הקובץ (ללא טעויות בדרך) לקחה פחות זמן מהצלחה סופית (success eventual) כאשר המשתמש טעה לפחות פעם אחת בדרך אך לבסוף הצליח לאתר את הקובץ המבוקש) והצלחה סופית לקחה פחות זמן מאחזור שנגמר בכישלון (המשתמש הודיע שאינו מסוגל לאתר את הקובץ כלל). במאמר זה, מחשב, קובץ וניווט הם תנאי המחקר שתחתם נבחנו הקשרים השונים בין המשתנים האחרים (הצד השמאלי של תרשים 2).

Paper 5 **Paper 4**

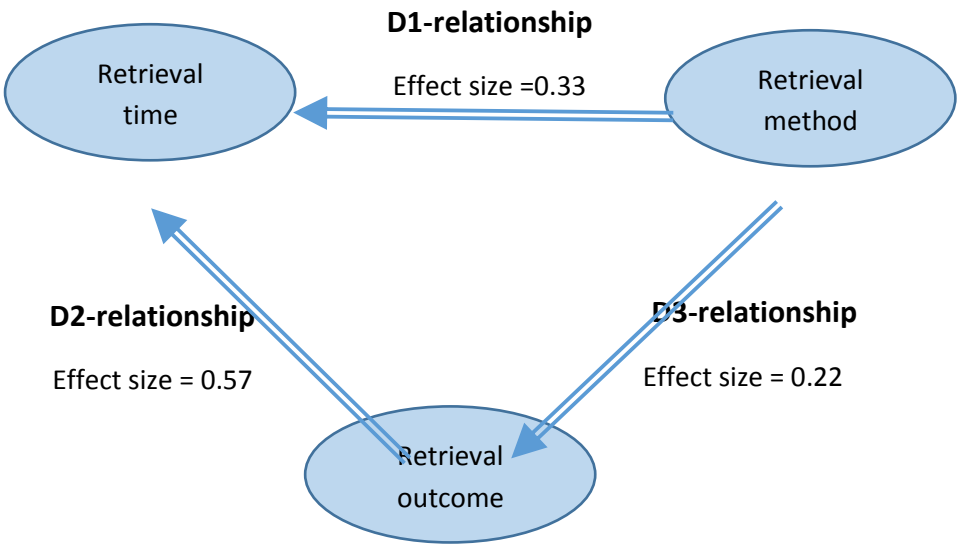




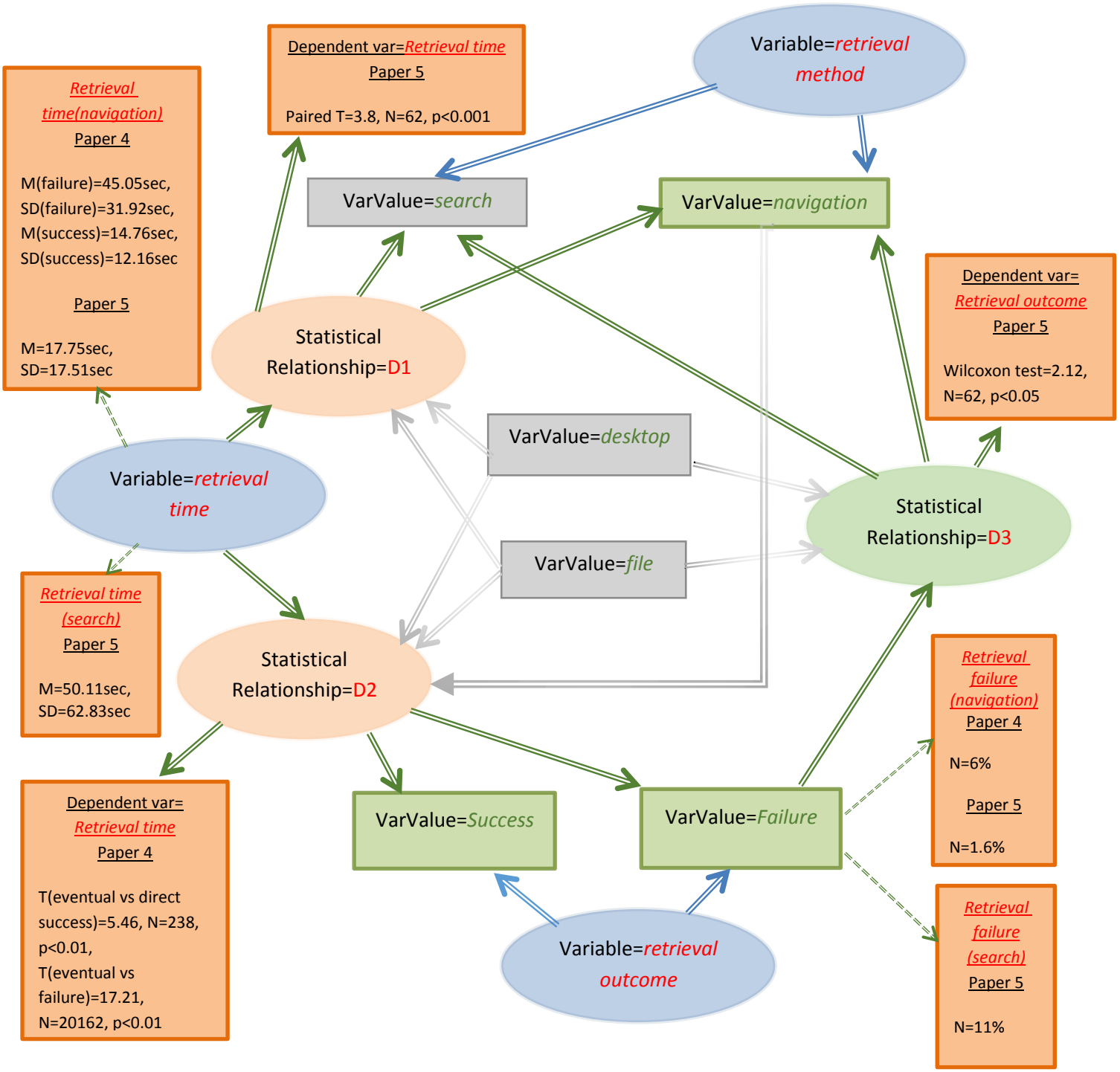


תרשים 2: גרפים של קשרים סטטיסטיים בין משתנים מתוך מאמרים 4, 5. הצמתים (האליפסות הכחולות) בדיאגרמה מייצגות משתנים סטטיסטיים והאליפסות הכתומות מייצגות קשרים סטטיסטיים. המלבנים הירוקים מייצגים את הערכים השונים של המשתנים. החצים הכחולים מייצגים את הקשר בין משתנה לערכיו, החצים הכתומים מייצגים את הקשר בין קשרים הסטטיסטיים למשתנים וערכיהם המקיימים קשר זה ולתוצאות הסטטיסטיות של הקשר. החצים הירוקים מקשרים בין הקשרים הסטטיסטיים לתנאים בהם הם נמדדו במחקרים. באופן עקרוני כל משתנה וערך של משתנה יכול להופיע פעם בתפקיד של משתנה תלוי, פעם בתפקיד של המשתנה הבלתי תלוי ופעם אחרת בתפקיד התנאי לקשר הנמדד בין זוג משתנים.

בשלב הראשון של בניית המודל הרשתי ניתנו שמות סטנדרטיים ייחודיים לכל משתנה וערכיו בשני המאמרים. לאחר מכן כל הופעה של זוג משתנים וקשר סטטיסטי ביניהם נותחה לחוד ונבנה גרף נפרד עבורה כפי שמודגם בתרשים 2. בתרשים זה עדיין לא ניתן לראות את התמונה המלאה והכללית המשתקפת משני המחקרים אלא כל קשר סטטיסטי שנמדד לגופו. בשלב הבא מיזגנו את הגרפים הללו לרשת אחת שבה כל משתנה מופיע פעם אחת בלבד והמידע על הקשרים הסטטיסטיים השונים אשר מגיע ממאמרים שונים משולב יחד לכדי תמונה אחת. תרשים 3 מציג את הרשת הדו-ממדית המציגה תמונה כללית של הקשרים בין המשתנים, ותרשים 4 מציג את הרשת התלת-ממדית המציגה את המידע והמבנה המפורט של הרשת.



תרשים 3: דיאגרמה מאוחדת דו-ממדית של משתנים וקשרים סטטיסטיים ביניהם מתוך מאמרים 4,5.



תרשים 4: דיאגרמה מאוחדת תלת-ממדית של משתנים וקשרים סטטיסטיים ביניהם מתוך מאמרים 4,5. הצמתים (האליפסות הכחולות) בדיאגרמה מייצגות משתנים סטטיסטיים והאליפסות הכתומות מייצגות קשרים סטטיסטיים. המלבנים הירוקים מייצגים את הערכים השונים של המשתנים והמלבנים הכתומים מייצגים את התוצאות הסטטיסטיים שהתקבלו במחקרים. החצים הכחולים מייצגים את הקשר בין משתנה לערכיו, החצים הכתומים מייצגים את הקשר בין קשרים הסטטיסטיים למשתנים וערכיהם המקיימים קשר זה ולתוצאות

הסטטיסטיות של הקשר. החצים הירוקים מקשרים בין הקשרים הסטטיסטיים לתנאים בהם הם נמדדו במחקרים.

לאחר שבנינו את הרשת המוצגת בתרשים 4 נוכל להשתמש בה לצורך מטא-אנליזה וכריית נתונים באופן הבא. ראשית, מתוך מבנה הרשת נצפה קשר ההשפעה המשולש בין שלושת המשתנים שנחקרו (זמן האחזור מושפע משיטת האחזור ותוצאתו ותוצאת האחזור מושפעת משיטת האחזור), כאשר כל מאמר השלים את ממצאי המאמר השני ותרם קשרים שונים לרשת עבור אותם המשתנים. מתוך הרשת ניתן לראות כי אחוז הכישלונות בניווט נמדד בשני המאמרים והיה נמוך יחסית, וכן ניתן לחשב אחוז כישלונות ממוצע לשני הניסויים שיעמוד על 3.8% (Fitchett & Cockburn, 2015). כמו כן, זמן אחזור בשיטת הניווט דווח גם הוא בשני המאמרים, אולם במאמר 4 נעשתה הבחנה בין זמן אחזור במקרי הצלחה ובין זמן האחזור במקרי כישלון, וכל אחד מן המקרים הללו חושב בנפרד, ואילו במאמר 5 לא נעשתה הבחנה זו. לכן, לשם השוואה בין תוצאות שני המאמרים ניתן לחשב עבור מאמר 4 את זמן האחזור הכללי כממוצע משוקלל כאשר ידוע שמספר הכישלונות עמד על 6% מן האחזורים (45.05*0.06 + 14.76*0.94=16.58) והוא 16.58 שניות בממוצע. תוצאה זו הינה קרובה מאוד לתוצאה שדווחה על עבור משתנים זה במאמר 5 והיא 17.21 שניות בממוצע לסך האחזורים. כלומר התובנה הנגזרת מניתוח הרשת עבור משתנה זה היא כי קיימת התאמה בין תוצאות הניסויים בשני המאמרים. כמובן שמדובר בדוגמא בלבד ולא ניתן לבצע הכללות בעזרת שני מאמרים בלבד, אולם במקרה זה ישנם מחקרים אחרים אשר השתמשו במתודולוגיות מחקר שונות המעידים על תוקף מתכנס ומאוששים את ממצאי המחקרים הללו (Bergman, Beyth-Marom, Nachmias, Gradovitch, & Whittaker, 2008; Fitchett & Cockburn, 2015).

בנוסף על בדיקת ההיתכנות, מחקר הפיילוט לימד אותנו גם על חלק מהקשיים שניתקל בהם בבניית המערכת ועל האופן שבו נעבור מתהליך ידני לשימוש בעיבוד שפה טבעית.

קשיים בהם נתקלנו בפיילוט

אחד הקשיים הבלתי צפויים בהם נתקלנו במהלך הפיילוט היה תנאים סמויים והומוגניים במאמרים מסוימים אשר הופכים להיות הטרוגניים במאמרים אחרים (או הטרוגניים בין מאמרים). לדוגמא במאמרים 4 ו-5 האחזורים נעשו במחשב בלבד ולכן הדבר לא צוין במפורש כתנאי במאמרים. כאשר נתקלנו במאמר המשווה אחזור במחשב לאחזור בטלפון נייד, היה עלינו לחזור למאמרים אלו ולהוסיף "מחשב" כתנאי בדיקה.

בנוסף, כפי שצוין בסקירת הספרות, משתנים שונים וערכיהם הופיעו בשמות שונים. לדוגמא, במאמר 5 הופיע המונח retrieval outcome וכן success rate כמתייחסים לאותו משתנה, ואילו במאמר 4 הופיע המונח percentages of success. במאמר 4 הופיע המונח retrieval method ואילו במאמר 5 הופיע המונח method for accessing files המתייחסים לאותו המשתנה. במאמר 5 המחקר בחן רק את שיטת הניווט (navigation) ולכן מונח זה הוחלף לעיתים קרובות עם המונח retrieval, כגון "navigation success" ו- "retrieval success" כבעלי אותה המשמעות. בעיה זו נפתרה כאמור בעזרת סטנדרטיזציה של שמות המשתנים. המערכת שבנינו לא מאפשרת להכניס נתונים לגבי משתנים טרם הוגדרו וקיבלו שם סטנדרטי.

המעבר מעבודה ידנית לשימוש בעיבוד שפה טבעית

על מנת לבחון כיצד יהיה ניתן לזהות את משתני המחקר והקשרים הסטטיסטיים ביניהם בצורה אוטומטית, זיהינו תבניות המכילות מלות מפתח אופייניות (כגון, t-test, test, Pearson ,t, r, p) שחוזרות על עצמן במאמרים, כגון, במשפטים (שמות המשתנים והקשרים מודגשים ורכיבים שחוזרים על עצמם צבועים בירוק):

"We found that the **number of words recalled** *after* **navigation** *was larger* **than the number of words recalled** *after* **search (*t*(26)=2.39**, p<0.05)"

"As we expected, **retrieval time** *for* **Direct Success navigations** *was shorter than for* **Eventual Success navigations** (***t*(1,060)=17.21**,p<0.01)".

"The question was tested by *comparing* **retrieval time, number of mistakes** *and* **percentage of failed retrievals**".

"*Comparison of* **retrieval time** *and* **number of mistakes** *used* a **paired t-test**…"

We tested the *relations* between **folder size** and **retrieval outcome** *using* **t-" "test**...

עבור משפטים לעיל ניתן להגדיר את שתי התבניות הבאות שמאפיינות אותן:

<[any word/s] [dependent variable name] [preposition] [independent variable value] [is/was] [adjective] than [independent variable value] [test name]=[result]>

<[comparison/comparing/relation] [preposition] [variable value name] [and] [variable value name] [use/using/used] [test name]>

לאחר זיהוי שמות המשתנים מתוך התבניות הבסיסיות ניתן לחפש את הופעותיהם של שמות אלו בטקסט ובכך לזהות תבניות נוספות. באמצעות תבניות חדשות ניתן לשלוף משתנים נוספים. תהליך זה יכול לחזור על עצמו מספר פעמים עד שלא ניתן למצוא משתנים חדשים.