מהפכת ה-Big Data מנקודת מבט של ארגוני הענק

ריאיון עם ד"ר יואל מארק, סמנכ"ל מחקר בחברת AMAZON

מראיינים: דודי סימן-טוב וסא"ל צ'

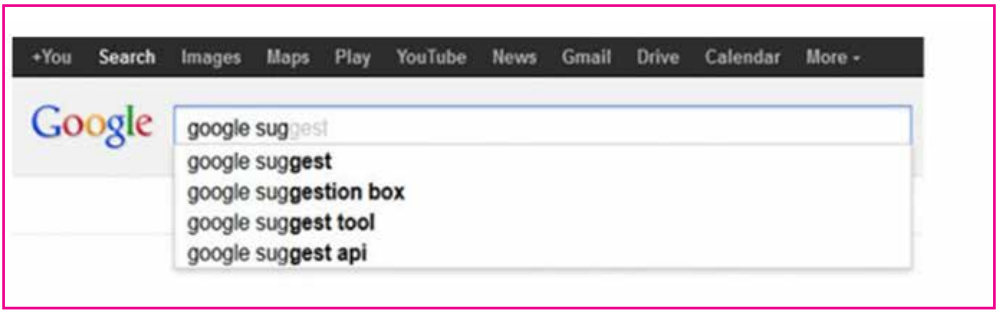
שאלה: ד"ר מארק, ספרי לנו בבקשה מה את חושבת על המונח "מהפכת ה־Big Data"?

תשובה: Big Data הוא מונח שיווקי.[1] בעבר השתמשנו במונח כריית מידע ("Data Mining"), וזה עדיין המונח המועדף בחוגים מקצועיים ואקדמיים. עם זאת, Big Data עוסק בחיבור של כמויות אדירות של מידע, עם משאבי מחשוב אדירים. החיבור הזה הוא בוודאי מהפכה.

שאלה: באיזה מובן זו מהפכה?

תשובה: כאשר אנחנו מדברים על Big Data, אנחנו צריכים לדבר על שני הצדדים של התופעה, החיובי והשלילי. הצד החיובי הוא ההזדמנויות הגדולות שיש לנו כעת.

באופן כללי, מנועי חיפוש בעולם נכשלים כתוצאה משלוש סיבות מרכזיות: או שאלגוריתם הדירוג (ranking) שלהם אינו טוב, או שהתוכן שהמשתמש מחפש אינו במאגר, או שהתוכן נמצא, אבל המשתמש אינו יודע לתרגם את הצורך שלו במידע לשאילתת חיפוש אפקטיבית. בניסיון לפתור את הבעיה השלישית, כשהייתי ב־IBM, פיתחנו, הקבוצה שלי ואני, אלגוריתמים לסיוע בשאילתות על פאלם פיילוט, אבל השתמשנו במסמכים עצמם כדי לייצר את השאילתות המוצעות, והשאילתות לא נראו טבעיות. מאוחר יותר, כאשר הצטרפתי ל־Google, גיליתי שיש אב טיפוס לסיוע בשאילתה, שיוצר על קבוצה סטטית של מסמכים, שרצה ב־labs.google.com. היופי בכלי היה שהוא השתמש כבסיס בשאילתות אמיתיות של משתמשים אמיתיים. קפצתי על ההזדמנות, והאתגר היה להפוך את הפונקציונליות הזו מסטטית לדינמית כך שתוכל לרוץ על כל השאילתות



1 מקור המונח הוא, ככל הנראה, ג'ון משי (John Mashey), המדען הראשי של חברת סיליקון גרפיקס בשנות ה־90 של המאה ה־20.

במנוע החיפוש של Google ולהתעדכן כל הזמן. אחרי שנתיים של עבודה בצוות הפיתוח בחיפה, השקנו את Google Suggest. הפרויקט הזה חייב כמויות עצומות של מידע (מיליוני משתמשים), ופלטפורמות חישוב מקבילות (MapReduce).

היום, באמזון, אנחנו לוקחים את האתגר צעד אחד קדימה. המשימה שלנו היא לעשות את מה שהכי טוב עבור הלקוחות שלנו, ואנחנו צריכים לעבד סיגנלים רבים, כדי להבין אותם ולספק את צורכיהם. זו הסיבה שאנחנו צריכים שירותי ענן כמו AWS (Amazon Web Services) שלא רק תומכים בעיבוד מקבילי אלא גם מקדמים שירותי עיבוד של למידת מכונה. בארגון שאני נמצאת בו, Alexa Shopping, אנחנו משתמשים ב־Big Data כדי להפוך את הרעיון של אינטליגנציה מלאכותית למציאות. אנחנו מאפשרים ללקוחות לשאול לגבי מוצרים ולרכוש אותם באמצעות דיבור למכונת Echo.[2] זה משנה את הפרדיגמה של התקשורת של בני אדם עם מכשירים, ונותן לנו מוטיבציה להרחיב את גבולות המדע. במרחב משבש המציאות הזה, לא היינו יכולים להתקדם ללא שימוש בכמויות עצומות של מידע ופלטפורמות חישוב חזקות, אבל גם בהרבה מומחי תוכן, כדי להבין באמת את הלקוחות ולהתחיל לגזור לאחור משם.

שאלה: האם זה משהו שארגוני מודיעין יכולים לעשות?

תשובה: הדוגמאות שהבאתי מבליטות את המגבלות שעומדות בפני ארגוני ביטחון ומודיעין כשהם מנסים להשתמש בטכניקות דומות. לארגונים ממשלתיים אין מיליוני משתמשים. הם צריכים לפצות על זה על ידי בקשה מהמשתמשים שלהם לייצר פידבק מפורש, שימוש בטכניקות למידת מכונה, ושימוש במומחים (domain experts) כדי לתייג מידע עבור האלגוריתם.[3] זה מאתגר, כי למידת מכונה פחות מוצלחת כשהיא פועלת על כמויות מידע קטנות, וארגוני מודיעין צריכים להמציא גישות חדשות, או להתחבר למשאבים חיצוניים, כדי לפצות על הכמויות הנמוכות יחסית של מידע פנימי (שמגיע ממשתמשים).

שאלה: ומה הצד השלילי של Big Data?

הצד השלילי של ה־ביג דאטה הוא כמובן הרבה זבל (Big Garbage). אלגוריתמים מתוחכמים, במיוחד אלגוריתמים של למידה עמוקה, מראים תוצאות מצוינות, אבל קשה מאוד להבין את פעולת המכונה, ולהסביר אותה. האלגוריתם הופך להיות מעין "קופסה שחורה" ואנחנו צריכים לסמוך על כך שהיא מבצעת את תפקידה היטב. לאחרונה יש עיסוק ציבורי בסכנות של אינטליגנציה מלאכותית, אבל אינטליגנציה מלאכותית מסוכנת רק כשמשתמשים באלגוריתמים טיפשיים ובמדענים טיפשים. אני מאמינה בגישה זהירה של בחינה מדוקדקת של האלגוריתם, והבנה מדוע אנחנו מקבלים תוצאות כאלה או אחרות. אנחנו קוראים לזה בר פירוש. אין זה אחראי למדען להגיד שהסיבה שהוא מקבל תוצאות מסוימות היא בגלל ש"כך המכונה החליטה". כל צעד צריך להיות מנוטר כך שהאנליסטים (מומחי התוכן) יוכלו לוודא את התוצאות. כמובן, שזה עוד יותר חשוב במקרה של ארגוני ביטחון ומודיעין, שמקבלים החלטות על חיים ומוות באמצעות

2 הרמקול/מיקרופון החכם של חברת Amazon, שנקרא Echo, שמחובר לשירות Alexa.

3 במובן מסוים, זהו חידוש של פרקטיקה ישנה, אך מסיבות שונות. בעבר, משתמשים היו מסמנים מילות מפתח בטקסטים, כי מנועי החיפוש היו מאנדקסים רק את מילות המפתח הללו. כיום, מומחים צריכים לסמן מילות מפתח, כדי לסייע לאלגוריתם למידת המכונה.

אלגוריתמים. יש עדיין הרבה מקום לפעולה של בני אדם, וזה נכון במיוחד משום שדוגמאות חיוביות עשויות להשתנות לאורך זמן.[4] אנליסטים צריכים להתאים את עצמם לעידן החדש הזה – להבין את הקונספט של מאפיינים, ולהבין איך הפעולות שלהם תורמות לפעולתה של המכונה הלומדת. ארגוני מודיעין צריכים לבנות לצורך העניין ספקטרום של הכשרות. חוקרים ואנליסטים צריכים לדעת לכתוב קוד, ומתכנתים צריכים להבין את התחום העסקי שבו הם פועלים. ברמה הבסיסית ביותר, ארגוני מודיעין צריכים לפחות לייצר קהילות הטרוגניות של מומחים עסקיים ומומחי טכנולוגיה, כדי לגשר על הפער.

שאלה: האם יש היבטים נוספים שבהם ארגוני מודיעין שונים מארגונים אזרחיים בהקשר של Big Data?

תשובה: ארגוני מודיעין עומדים בפני אתגר קשה יותר מאשר ארגונים מסחריים אזרחיים, גם משום שהם סגורים, ולעיתים קרובות מוגבלים ביכולת שלהם להשתמש בגישות של ניסוי וטעייה. הם רגישים מאוד לתוצאות של כיסוי (recall) ולא רק לדיוק (precision).[5] הבעיה שלהם היא לפיכך מסובכת יותר. בניגוד לעוזרים אלקטרוניים, שהרבה אנשים מוכנים להשתמש בהם, אף שהם עדיין לא מושלמים, וכך לספק מידע שימושי שמאפשר להמשיך לשפר את המוצר, משתמשים מודיעיניים הם פחות סבלניים.

ארגוני מודיעין עומדים בפני קושי משמעותי נוסף. בגלל שהם סגורים, הם מוגבלים ביכולתם להשתמש בתשתיות ענן ושירותי תוכנה ציבוריים. חברת Amazon, למשל, מפתחת את העוזר האלקטרוני Alexa כפלטפורמה, ומאפשרת לארגונים אחרים לבנות תחומי ידע נקודתיים עבור Alexa ("Skills") מעל הפלטפורמה. זה מסייע גם לארגונים הללו, שאינם צריכים לפתח בעצמם אפליקציית זיהוי דיבור מאפס, וגם מסייע לחברת Amazon להוסיף פונקציונליות נוספת עבור לקוחותיה. פחות ופחות ארגוני תוכנה בעולם בונים את השירותים שלהם עבור ארגונים "סגורים", וארגוני מודיעין עלולים למצוא את עצמם נעולים וללא גישה לעושר שירותי התוכנה והמשאבים שקיימים במרחב הציבורי.

שאלה: מה לדעתך תהיה קפיצת המדרגה הבאה?

תשובה: אני מאמינה שמכונות שמופעלות על ידי קול הן המהפכה הבאה. בעתיד, יופעלו כל המכונות על ידי קול. אשאיר לארגוני המודיעין לנתח את ההזדמנויות שעומדות בפניהם בנושא.

---

4 Positive examples הן דוגמאות שהמכונה משתמשת בהן כדי ללמוד. הכוונה כאן היא למקרים שבהם יש השתנות בתופעה עצמה, ואז צריך לייצר למידה מחודשת.

5 המונח recall מתאר עד כמה מנוע חיפוש מוצא את כל התוצאות הנכונות האפשריות. המונח precision מתאר עד כמה כל תוצאה שמחזיר מנוע החיפוש היא אכן נכונה. אלה שני המדדים הנפוצים להשוואה בין מנועי חיפוש, ומתקיים ביניהם יחס הפוך (ככל שמשפרים את הדיוק פוגעים בכיסוי ולהפך), אך לא ליניארי. כך, למשל, מנוע החיפוש של Google, הוא מדויק מאוד, ולעיתים קרובות אנשים מקבלים תשובה לשאלתם כבר בתוצאות הראשונות שהוא מחזיר, ויש בו דגש נמוך בהרבה על כיסוי. ייתכן שיש הרבה תוצאות מאוד רלוונטיות למשתמש, שלא חוזרות כלל בתוצאת החיפוש, והמשתמש לא יודע עליהן. במקרים שבהם החמצה של ידיעה חשובה עלולה לעלות בחיי אדם, בחירה בזו במדדי recall/precision היא מסובכת.