זיהוי וחיזוי אנומליות בטוויטר באמצעות נתונים גיאוגרפיים

לירון מרקוס

בהנחיית: ד"ר רון בקרמן

שירותי מדיה חברתיים כגון טוויטר מייצרים כמות הולכת וגדלה של מידע על אירועים יומיומיים. היכולת לנטר מידע בקנה מידה גדול זה מאפשר גילויים כמעט בזמן אמת של אירועים הנעים מפסטיבלים מקומיים לאסונות טבע ברמה ארצית. בכך הופך טוויטר לכלי חישה (Sakaki et al. 2010) אשר מאפשר לחוקרים לענות על מגוון שונה של שאלות כמו כיצד מידע ברשת חברתית מופץ בין משתמשים בהקשר מקומי והיבטים תרבותיים.
טוויטר פיתחה כלי תיוג גיאוגרפי המאפשר לתייג טוויט עם המיקום הנוכחי של המשתמש, אולם האתגר העיקרי הינו לחלץ מיקומים מטוויט אשר אינם מכילים את התיוג הגיאוגרפי.
בעבודה זו, אנו מציגים גישה גרפית חדשנית המבוססת על טכניקת מונחה למחצה (semi-supervised) לגילוי מקומות על בסיס תוכנו של הטוויט. שלא כמו בגישות קיימות, אשר מתמקדות בחילוץ מילת מפתח גיאוגרפית (ברמות צפיפות של מדינה/פרובינציה/מחוז/עיר) או בניסיונות לגילוי מהיכן הטוויט בוצע, אנו שואפים לגילוי מיקומים אשר הטוויט מצביע ברמת צפיפות גבוהה עד כדי רמת רחוב או איזור עניין (POI).
אנו בונים גרף משולב של משתמשי טוויטר ואיזכורים גיאוגרפיים ומתגברים על דו-משמעות של מיקומים בעלי שמות גיאוגרפיים ע"י ניצול התכונות המקומיות של הגרף: בהינתן סט של מיקומים בעלי שמות חד-משמעיים, אנו קושרים את המיקומים בעלי דו-משמעות הנמצאים בסמיכות למיקומים בעלי שמות חד-משמעיים בגרף. לשם כך, אני מציעים מודל הסתברותי חדש לשיטת הפצת תגים (Label Propagation) בגרף ומפיקים אלגוריתם מקבילי לתיוג אופטימלי של צמתים בגרף בעל קנה-מידה גדול. אנו קוראים לתוצאת השיטה החדשה של הפצת התגים – Splash Label Propagation (SLP).