

Detekcia a sledovanie spravodajských tém na Webe inšpirované sociálnym hmyzom

Štefan Sabo

školiteľ: prof. Pavol Návrat

Prehľad

- Metóda na dynamickú identifikáciu a sledovanie spravodajských príbehov, ktoré sa momentálne odvíjajú na Webe.
- Používame skupinu agentov inšpirovaných správaním sa včiel medonosných pri zbere potravy v prírode.
- Súvislosti medzi článkami vyhodnocujeme postupne, na základe jednotlivých príbehových slov, pričom najslubnejšie príbehové slová, ktoré umožňujú identifikovať prepojenia medzi článkami sú propagované, menej vhodné slová sú opúšťané.
- Výsledkom činnosti agentov je graf súvislostí medzi článkami, ktorý následne analyzujeme za účelom získania jednotlivých spravodajských príbehov.

Motivácia

- Dynamická identifikácia príbehov umožňuje identifikovať spravodajské príbehy a sledovať ich zároveň s tým, ako sa odvíjajú.
- Identifikácia príbehu je výzvou, keďže je náročné definovať, čo spravodajský príbeh predstavuje.
- Využiteľné pre záujemcu o aktuálne dianie, ale aj pre potreby spravodajskej analýzy.

Výzvy

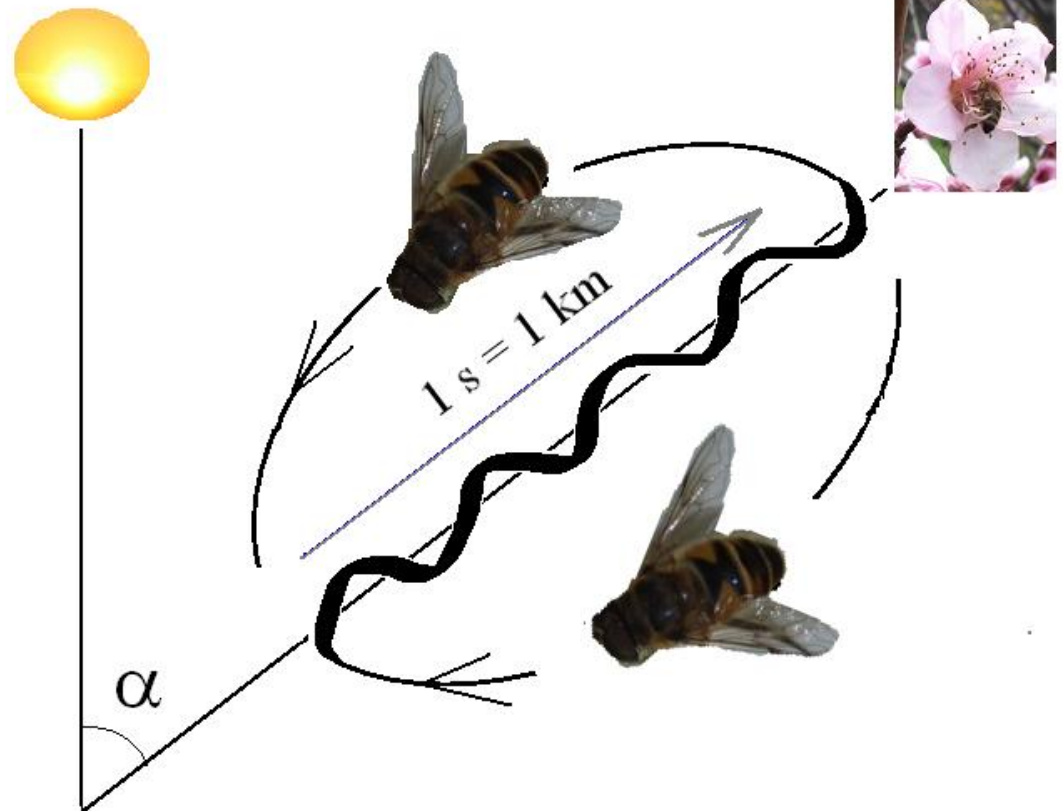
- Problematickou oblasťou je reprezentácia príbehu vhodná pre dynamické sledovanie.
- Bežne využívané postupy detekcie a sledovania tém (TDT) sa spoliehajú na globálu analýzu článkov, napríklad *term frequency – inverse document frequency*.
- Korpus všetkých článkov nie je za behu k dispozícii.
- Na zaznamenávanie nových aktuálnych článkov a sledovanie zmien v článkoch starších je potrebný dynamický systém.

Inšpirácia sociálnym hmyzom

- Populárny prístup k riešeniu optimalizačných / vyhľadávacích problémov.
- Prevažne stochastické systémy.
- Dôraz na jednoduchosť agentov, komplexita systému vyplýva z interakcií, nie zo správania jednotlivcov.
- Okrem včiel využívané aj mravce, termity, svätojánske mušky a mnohé iné.

Včelí tanec

- decentralizovaný
- koordinovaný na systémovej úrovni
- umožňuje dynamickú reakciu



Príbehové slová

- Spravodajské príbehy je často možné reprezentovať jednoduchými výrazmi, ktoré vystihujú podstatu príbehu, alebo slúžia ako krátke popis príbehu.
- Tieto príbehové slová sú zaujímavé z pohľadu navigácie v priestore spravodajských príbehov.
- Slúžia nám ako zdroje, ktoré je individuálne možné jednoducho sledovať pomocou agentov.

Metóda – získavanie príbehových slov

- Príbehové slovo predstavuje výraz relevantný k určitej skupine článkov, ktorý tieto články prepája na základe ich relevancie k spoločnému spravodajskému príbehu.
- Namiesto priamej identifikácie príbehov relevantných k danému článku sa snažíme pre daný článok identifikovať len skupinu relevantných príbehových slov.
- Výhoda tohto prístupu spočíva v tom, že určiť relevanciu dvoch článkov vzhľadom na zvolené príbehové slovo môžeme aj bez znalosti širšieho kontextu.
- Príbehy získavame ako výsledok emergentného správania sa roja agentov.

Metóda – získavanie príbehových slov

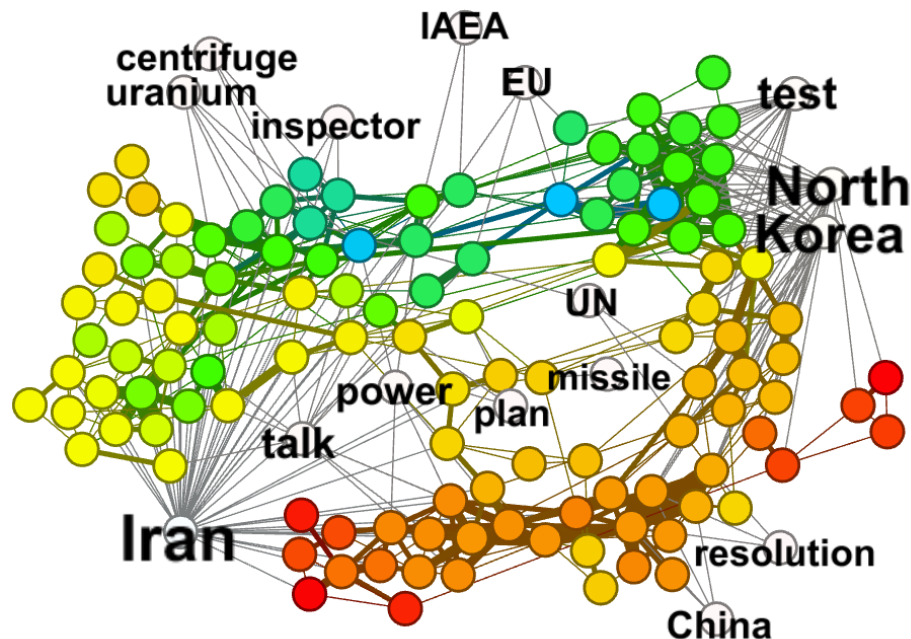
- Agenti sa presúvajú medzi článkami a snažia sa v nich identifikovať význačné príbehové slová.
- Ak je jedno príbehové slovo súčasne relevantné k viacerým článkom, medzi týmito článkami vzniká väzba na základe tohto príbehového slova.
- Porovnanie všetkých možných dvojíc článkov vzhľadom na každé potenciálne príbehové slovo by bolo nepraktické, preto využívame stratégiu výberu založenú na správaní sa včiel medonosných.

Stratégia výberu príbehových slov

- Každý agent nesie nanajvýš jedno príbehové slovo
- Tri činnosti, ktorým sa agenti môžu venovať:
 - Zber potravy
 - Včelí tanec
 - Sledovanie
- Pri zbere potravy agenti príbehové slová vyhodnocujú.
- Počas včelieho tanca agenti príbehové slová propagujú.
- Sledovanie agenti zahajujú v prípade, že opustili svoje príbehové slovo a chcú si vybrať nové.

Metóda – identifikácia príbehu

- Využíva grafovú reprezentáciu získaných údajov, pričom články a príbehové slová predstavujú uzly a relácie relevancie predstavujú hrany.

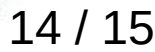


Metóda – identifikácia príbehu

- Graf je dynamicky konštruovaný agentmi počas toho, ako sú články skúmané.
- Články sú zoskupené do príbehov pomocou Louvainovho algoritmu na detekciu komunít.
- Po prvotnej identifikácii príbehov je možné nové články klasifikovať za behu bez nutnosti prepočítavať komunity od základu.
- Detekciu komunít je potrebné vykonávať jedine v prípade, že chceme aktualizovať celú sadu príbehov.

Výsledky – presnosť identifikácie

príbeh	príbehových slov	článkov	presnosť [%]
Bombové útoky v Bostone	18	64	96.88
Investičné správy	4	16	93.75
Cyperský dlh	10	56	89.29
Zrážky v Egypte	2	9	88.89
Bank of Japan	10	26	88.46
Voľby v Taliansku	11	37	83.78
Dlhová kríza	5	24	83.33
Napätie v Severnej Kórei	13	71	74.65
Voľby v Iraku	4	12	58.33
Zbrojné zákony v USA	11	49	55.10
Zemetrasenie v Iráne	12	37	18.92
Spolu	100	401	75.56



Zhrnutie

- Navrhovaný prístup využíva agentov na identifikáciu príbehových slov pre skupinu spravodajských článkov.
- Články sú porovnávané a ich príbuznosť je vyhodnocovaná na základe rôznych príbehových slov, pomocou stratégie založenej na správaní sa včiel medonosných.
- Dynamická povaha tohto procesu umožňuje agentom flexibilne reagovať na nové články, alebo na zmeny v článkoch už spracovaných.
- Príbehy sú identifikované za behu na základe identifikácie komunít v grafe článkov.
- Tento prístup nevyžaduje natrénovanie agentov, alebo globálnu analýzu korpusu.