



Дәріс 14: Әлеуметтік желілерді талдау және ұсыныс жүйелері

Оқытушы: Сарсембаева Т.С.

Дәрістің мақсаты

Заманауи цифрлық әлемнің байланысқан құрылымын және оны талдаудың екі негізгі технологиясын түсіндіру: Әлеуметтік желілерді талдау (SNA) және Ұсыныс жүйелері (RecSys).

Осы мақсат аясында, біріншіден, SNA-ның негізі болып табылатын Графтар теориясын және орталықтықты өлшеу әдістерін меңгерту көзделеді. Бұл әдістерге Аралық орталықтықты (Betweenness Centrality) желідегі "көпір" немесе "делдал" ретінде түсіндіру, сондай-ақ Меншікті векторлық орталықтықтың (Eigenvector Centrality) "маңызды адамдармен байланысты болу" принципін талдау кіреді.

Сонымен қатар, PageRank алгоритмін, оның "кездейсоқ серфер" моделін және Демпинг факторын (d) түсіну қарастырылады. Екіншіден, дәріс Ұсыныс жүйелерінің негізгі міндетін және қарапайым әдістерін зерттеуге бағытталған.

Негізгі сұрақтар

1. Әлеуметтік желілерді талдау (SNA) мен Ұсыныс жүйелері (RecSys) арасында қандай айырмашылық бар?
2. Граф дегеніміз не? "Түйін" (Node) мен "Қабырға" (Edge) ұғымдарына мысал келтіріңіз.
3. Аралық орталықтық (Betweenness Centrality) неліктен "делдалдық" деп аталады?
4. PageRank меншікті векторлық орталықтықтан несімен ерекшеленеді (Дауысты бөлу)?
5. Ұсыныс жүйелерінде "Пайдаланушы-Нысан матрицасы" не үшін қолданылады және ол неліктен "сирек" (sparse) болып келеді?
6. Бірлескен сүзгілеудегі (Collaborative Filtering) "Пайдаланушыға негізделген" және "Нысанға негізделген" әдістердің айырмашылығы неде?
7. "Салқын бастау" (Cold Start) мәселесі дегеніміз не және ол қашан туындайды?

Байланысқан әлемді талдау

Біздің заманауи цифрлық өміріміз толығымен желілерге негізделген. X (Twitter) және LinkedIn желілеріндегі байланыстар, Netflix пен YouTube-тегі видео ұсыныстары, Amazon-дағы тауарлар – осының барлығының негізінде екі іргелі технология жатыр:



Әлеуметтік желілерді талдау (SNA)

Адамдар мен идеялардың құрылымын және ықпалын түсіну. "Бұл желідегі ең маңызды тұлға кім?", "Ақпарат қалай таралады?"



Ұсыныс жүйелері (RecSys)

Сол желідегі мінез-құлықты талдау арқылы пайдаланушының қалауларын болжау. "Бұл адамға келесіде не ұнайды?"

Бұл екі сала бір-бірімен тығыз байланысты және заманауи интернеттің қозғалтқышы болып табылады.

Әлеуметтік желілерді талдау (SNA)

SNA – бұл әлеуметтік құрылымдарды зерттеу үшін графтар теориясын қолданатын пән. Әлеуметтік желі – бұл математикалық тұрғыдан граф $G = (V, E)$ ретінде сипатталатын құрылым:

Түйіндер (Nodes/Vertices, V)

Желідегі нысандар. Мысалы: X-тегі пайдаланушылар, ғалымдар, веб-парақшалар.

Қабырғалар (Edges/Links, E)

Түйіндер арасындағы байланыс. Мысалы: "достық" (Facebook), "оқу" (following, X), "сілтеме" (Web).

Бағытталған (Directed) граф: Байланыстың бағыты маңызды (мысалы, X-те A пайдаланушысы B-ны оқиды, бірақ B A-ны оқымауы мүмкін).

Бағытталмаған (Undirected) граф: Байланыс екіжақты (мысалы, LinkedIn-дегі "байланыс").

SNA-дағы басты сұрақ: "Осы графтың ішіндегі ең маңызды немесе орталық түйін қайсысы?" Бірақ "маңыздылық" әртүрлі болуы мүмкін. Оны өлшеу үшін орталықтық (centrality) өлшемдері қолданылады.

Аралық орталықтық (Betweenness Centrality)

Бұл сіздің сұранысыңыздағы "Медиация орталығы" ұғымына жақын, яғни "делдалдық" немесе "аралықта болу" дегенді білдіреді.

Түйіннің маңыздылығы оның желідегі ақпарат ағынында "көпір" немесе "делдал" ретінде қызмет етуімен өлшенеді. Егер түйін көптеген басқа түйіндердің арасындағы байланыс жолында жатса, оның маңыздылығы жоғары.

Техникалық анықтама: Бұл түйін желідегі барлық басқа түйін жұптарының арасындағы ең қысқа жолдарда (shortest paths) қаншалықты жиі орналасатынын есептейді.

Формула (Концептуалды):

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Мұнда:

- σ_{st} – s түйінінен t түйініне дейінгі барлық ең қысқа жолдардың жалпы саны.
- $\sigma_{st}(v)$ – s-тан t-ға дейінгі сол ең қысқа жолдардың u түйіні арқылы өтетін саны.

Меншікті векторлық орталықтық (Eigenvector Centrality)

Маңыздылық тек байланыс санымен өлшенбейді. Сіздің маңыздылығыңыз – сіз байланысқан адамдардың маңыздылығына тәуелді.

Техникалық анықтама: Бұл рекурсивті анықтама. u түйінінің орталықтығы (x_u) оның барлық көршілерінің ($N(u)$) орталықтықтарының (x_t) қосындысына пропорционал.

$$x_u = \frac{1}{\lambda} \sum_{t \in N(u)} x_t$$

Мұнда λ – тұрақты сан (ең үлкен меншікті мән). Бұл есеп $Ax = \lambda x$ түріндегі меншікті вектор мәселесін шешуге келіп тіреледі, мұнда A – графтың іргелестік матрицасы (adjacency matrix).

PageRank

PageRank – бұл Google-дың негізін қалаушылары Ларри Пейдж бен Сергей Брин әзірлеген, меншікті векторлық орталықтықтың арнайы және өте танымал бір түрі. Ол бастапқыда веб-парақшаларды рангтеу (дәрежелеу) үшін бағытталған графтарға (directed graphs) арналған.

PageRank-тің негізгі принциптері

1 "Дауыс беру"

А парақшасынан В парақшасына сілтеме – бұл А-ның В үшін берген "дауысы".

2 Дауыстың салмағы (Рекурсия)

Жоғары PageRank-і бар (маңызды) парақшадан келген дауыстың салмағы, төмен PageRank-і бар парақшадан келген дауыстан әлдеқайда жоғары. Бұл – меншікті векторлық орталықтықпен ұқсас жері.

3 Дауысты бөлу (Маңызды айырмашылық)

Егер маңызды А парақшасы 10 басқа парақшаға сілтеме жасаса, ол өзінің "маңыздылық дауысын" сол 10 парақшаға теңдей бөліп береді. Егер ол тек бір ғана В парақшасына сілтеме жасаса, В парақшасы оның барлық дауысын алады.

"Кездейсоқ серфер" моделі (Random Surfer Model)

PageRank-ті түсінудің ең оңай жолы: Интернеттегі кездейсоқ пайдаланушыны елестетіңіз. Ол бір парақшадан екіншісіне сілтемелер арқылы тоқтаусыз секіреді. PageRank – бұл осы "кездейсоқ серфердің" белгілі бір уақыт мезетінде кез келген парақшада болуының ықтималдығы. Көптеген маңызды сілтемелер алатын парақшаларға серфер жиі түседі.

"Тұйыққа тірелу" мәселесі және Демпинг факторы (d)

Егер серфер сілтемесі жоқ парақшаға (тұйық) түссе немесе тек бір-біріне сілтеме жасайтын парақшалар тобына (цикл) түсіп қалса не болады? Ол сол жерде "мәңгі" қалып қояды.

Бұл жерде **Демпинг факторын (d)** пайдаланамыз, әдетте $d = 0.85$.

Серфер 85% ықтималдықпен парақшадағы кездейсоқ сілтемені басады, ал 15% ықтималдықпен ол "жалығады" да, интернеттегі кез келген басқа парақшаға "телепорт жасайды".

Бұл әдіс тұйықтардан шығуға кепілдік береді және әрбір парақшаның минималды PageRank мәні болуын қамтамасыз етеді.

Ұсыныс жүйелері (Recommender Systems)

Ұсыныс жүйелері – бұл пайдаланушылардың қызығушылықтарын болжау арқылы оларға ең сәйкес келетін нысандарды (items) (мысалы, фильмдер, тауарлар, музыка) ұсынатын алгоритмдер.

Негізгі мәселе: Пайдаланушы-Нысан матрицасы (User-Item Matrix)

Барлық ұсыныс жүйелерінің негізінде осы үлкен матрица жатыр.

- Жолдар = Пайдаланушылар (мысалы, 10 миллион)
- Бағандар = Нысандар (мысалы, 1 миллион фильм)
- Ұяшықтар = Пайдаланушының нысанға қойған бағасы (мысалы, 1-ден 5-ке дейін) немесе қарағаны/қарамағаны (0/1).

Бұл матрица әрқашан өте сирек (sparse) болады, себебі бір адам барлық 1 миллион фильмді көре алмайды. Матрицаның 99.9%-дан астамы бос ("?", "0") болады. Ұсыныс жүйесінің міндеті – осы бос ұяшықтарды ("?", "0") болжау.

Қарапайым әдістер

Автоматты емес кураторлық

Ұсыныстарды арнайы сарапшылар (редакторлар, кураторлар) қолмен жасайды.
Мысалы: Spotify-дағы "Редактор таңдауы" плейлисті, Netflix-тің басты бетіндегі "Жаңа және танымал" (көбіне қолмен таңдалады), жаңалықтар сайтының "Басты жаңалық" бөлімі.

Танымалдылыққа негізделген ұсыныстар

Ең қарапайым автоматтандырылған әдіс. Барлық пайдаланушыға бірдей, ең танымал (көп қаралған, көп сатып алынған, жоғары бағаланған) нысандарды ұсыну. Мысалы: YouTube-тағы "Трендте" (Trending) бөлімі, Netflix-тегі "Қазақстандағы Top-10", X/Twitter-дегі "Аktуальды тақырыптар".

Бірлескен сүзгілеу (Collaborative Filtering, CF)

Бұл – ұсыныс жүйелеріндегі ең маңызды және кең таралған әдіс. "Маған мен сияқты басқа адамдарға ұнаған нәрселерді ұсын". Бұл әдіс нысанның мазмұнына (фильмнің жанры, кітаптың мәтіні) қарамайды. Ол тек Пайдаланушы-Нысан матрицасындағы мінез-құлыққа (бағалау, сатып алу) сүйенеді.

1. Пайдаланушыға негізделген CF (User-based CF)

1. Ұқсас пайдаланушыларды табу. Сізге (Алиса) ең ұқсас пайдаланушыны (мысалы, Бауыржан) табу. Ұқсастық олардың Пайдаланушы-Нысан матрицасындағы векторларының (яғни, олардың бағаларының) ұқсастығы арқылы есептеледі (мысалы, Косинустық ұқсастық немесе Пирсон корреляциясы).
2. Ұсыныс жасау. Бауыржан жоғары бағалаған (мысалы, "Inception" фильміне 5 жұлдыз), бірақ Алиса әлі көрмеген фильмдерді Алисаға ұсыну. Болжам Бауыржанның (және басқа да ұқсас пайдаланушылардың) бағаларының орташа салмақталған мәні ретінде есептеледі.

2. Нысанға негізделген CF (Item-based CF)

Бұл әдісті (Amazon танымал еткен) мәселені керісінше шешеді: пайдаланушыларды емес, нысандарды салыстырады.

1. Ұқсас нысандарды табу. Егер "X" фильмі мен "Y" фильміне барлық пайдаланушылар ұқсас баға берсе (яғни, X-ты ұнатқандар Y-ты да ұнатады, X-ты ұнатпағандар Y-ты да ұнатпайды), бұл екі фильм "ұқсас" болып саналады. Бұл ұқсастық матрицадағы бағандарды салыстыру арқылы есептеледі.
2. Ұсыныс жасау. Егер сіз "Бэтмен: Бастама" фильміне жоғары баға берсеңіз, жүйе сізге оған "ұқсас" (яғни, басқа пайдаланушылар да бірге жоғары бағалаған) "Қара сері" фильмін ұсынады.

"Салқын бастау" (Cold Start) мәселесі

Бірлескен сүзгілеудің (CF) ең үлкен әлсіздігі – **"салқын бастау" мәселесі**:

→ Жаңа пайдаланушы

Жүйеге жаңа пайдаланушы тіркелді. Оның ешқандай бағасы (тарихы) жоқ. Жүйе оған ұқсас пайдаланушыны да таба алмайды, оның не ұнататынын да білмейді.

→ Жаңа нысан

Платформаға жаңа фильм қосылды. Оны ешкім әлі бағаламады. CF бұл фильмді ешкімге ұсына алмайды, себебі оның басқа нысандарға "ұқсастығы" белгісіз.

Осы мәселелерді шешу үшін танымалдылыққа негізделген (жаңа қолданушыға Топ-10 көрсету) және келесі бөлімдегі әдістер қолданылады.

Дәрісті меңгеруге арналған бақылау сұрақтары

1. Әлеуметтік желілерді талдау (SNA) мен Ұсыныс жүйелері (RecSys) арасында қандай айырмашылық бар?
2. Граф дегеніміз не? "Түйін" (Node) мен "Қабырға" (Edge) ұғымдарына мысал келтіріңіз.
3. Аралық орталықтық (Betweenness Centrality) неліктен "делдалдық" деп аталады?
4. PageRank алгоритмінде "Демпинг факторы" (d) не үшін қажет?
5. Ұсыныс жүйелерінде "Пайдаланушы-Нысан матрицасы" не үшін қолданылады және ол неліктен "сирек" (sparse) болып келеді?
6. "Нысанға негізделген" Бірлескен сүзгілеу (Item-based CF) қалай жұмыс істейді?
7. "Салқын бастау" (Cold Start) мәселесі дегеніміз не және ол қашан туындайды?

Әдебиеттер тізімі

1. Easley, D., & Kleinberg, J. (2010). Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press.
<https://www.cs.cornell.edu/home/kleinber/networks-book/>
2. Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7), 107–117. [https://doi.org/10.1016/S0169-7552\(98\)00110-X](https://doi.org/10.1016/S0169-7552(98)00110-X)
3. Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommender algorithms. Proceedings of the 10th International Conference on World Wide Web (WWW'01). <https://doi.org/10.1145/371920.372071>
4. Ricci, F., Rokach, L., & Shapira, B. (Eds.). (2022). Recommender Systems Handbook (3rd ed.). Springer.
<https://link.springer.com/book/10.1007/978-1-0716-2197-4>
5. NetworkX Development Team. (2025). NetworkX Documentation. <https://networkx.org/documentation/stable/>
6. Hug, N. (2020). Surprise: A Python scikit for recommender systems. <http://surpriselib.com/>