

دانشگاه تربیت مدرس

دانشگاه تربیت مدرس

دانشکده فنی مهندسی

بخش مهندسی صنایع

سمینار کارشناسی ارشد

شناسایی بیماری توسط وب

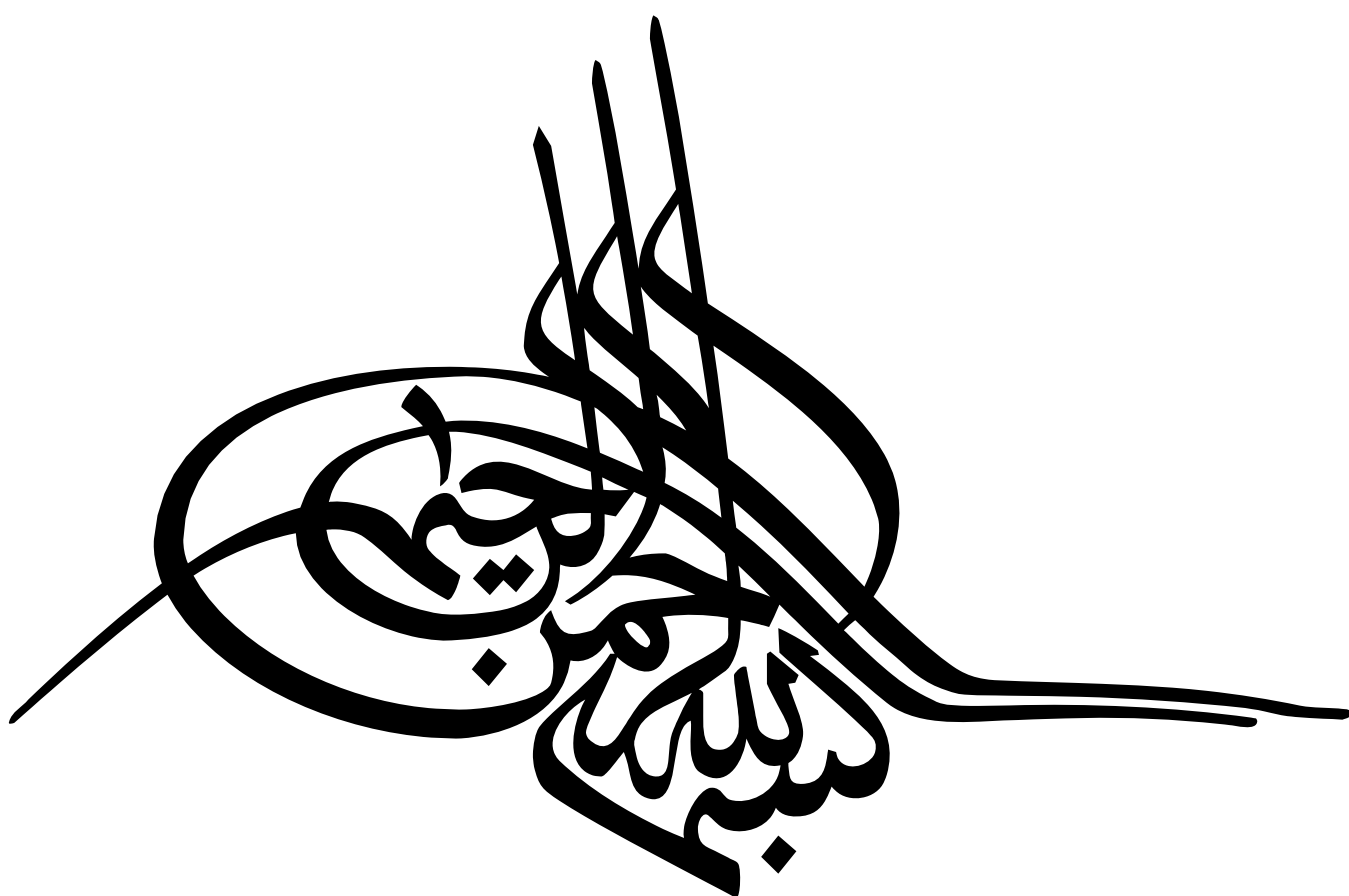
نگارش

سید علی لاجوردی

استاد

دکتر مهرداد کارگری

تیر ماه ۱۳۹۲



چکیده

با توسعه وب و ظهور web2.0 اطلاعات مختلفی توسط کاربران در محیط وب منتشر شد و با گسترش وبلاگ ها، انجمن های بحث و گفتگو، اخبار و خدمات آنلاین بخشی از محتوای وب سایت ها که مربوط به بهداشت و پزشکی است نیز در حال گسترش است.

تجزیه و تحلیل پیام های کاربران در رسانه های اجتماعی و وب می تواند ویژگی های جمعیت های مختلف، از جمله متغیر های بهداشت عمومی را اندازه گیری کند. به عنوان مثال در کار های اخیر نشان داده شده پیام های تویتر با نرخ آنفولانزا در ایالات متحده در ارتباط است.

در این سمینار به بررسی پتانسیل وب به عنوان یک ابزار مناسب برای اندازه گیری سلامت و نظارت پرداخته می شود.

در ادامه ابتدا به بررسی منابع مختلف اطلاعات سلامت در وب پرداخته می شود و سپس به مشکلات استخراج صحیح اطلاعات از آنها و مشکلات زبان فارسی اشاره شده و در نهایت با بررسی ابزارهای مختلف پیش بینی بر اساس داده کاوی در محیط وب (وب کاوی)، ابزارهای نظارتی در بهداشت و سلامت در محیط وب ارائه می شود.

کلیدواژه: وب کاوی، وب اجتماعی، تجزیه و تحلیل فناوری اجتماعی، نظارت بهداشتی، سلامت عمومی

فهرست

چکیده.....	أ
فهرست.....	ب
فصل اول - معرفی و کلیات.....	۱
۱-۱. مقدمه.....	۱
۲-۱. بیان مسئله و اهداف مطالعه.....	۲
۳-۱. اهمیت موضوع.....	۲
۴-۱. اهداف مطالعه.....	۳
۵-۱. تاریخچه موضوع.....	۳
۶-۱. تعاریف و اصطلاحات.....	۴
۷-۱. علائم و اختصارات.....	۵
۸-۱. ساختار کلی گزارش.....	۶
۹-۱. خلاصه فصل.....	۶
فصل دوم - منابع داده های سلامت در وب.....	۷
۱-۲. مقدمه.....	۷
۲-۲. شبکه های اجتماعی (Social Network).....	۸
۱-۲-۲. ویژگی های مربوط به رسانه های اجتماعی.....	۹
۲-۲-۲. طبقه بندی تعاملات اجتماعی شبکه بهداشت.....	۹
۳-۲-۲. حالت های ارتباطاتی با رسانه های اجتماعی.....	۱۰
۳-۲. میکرو بلاگ (Microblog).....	۱۰
۴-۲. وبلاگ ها و اخبار (News & Blogs).....	۱۲
۵-۲. تالارهای گفتمان (Forums).....	۱۳
۱-۵-۲. ویژگی های تالار های گفتمان بهداشت.....	۱۴
۶-۲. روند های جستجو (Google trends).....	۱۵
۷-۲. منابع کاربر محور.....	۱۶

۱۶.....	۲-۷-۱. چه کسی بیمار است: نمایش قدرت نقشه های تولید شده توسط کاربر
۱۶.....	۲-۷-۲. We Feel Fine : تشخیص متون آموزنده و عاطفی در وب اجتماعی
۱۶.....	۲-۵. خلاصه فصل
۱۸.....	فصل سوم - استخراج اطلاعات سلامت از وب
۱۸.....	۳-۱. مقدمه
۲۱.....	۳-۲. سیستم یکپارچه زبانی (UMLS)
۲۲.....	۳-۳. داده کاوی رسانه های اجتماعی (WSM)
۲۳.....	۳-۴. استخراج کننده اطلاعات (IE)
۲۳.....	۳-۴-۱. معیارهای کیفیت
۲۴.....	۳-۵. سیستم بازیابی اطلاعات وب (WebIRS)
۲۶.....	۳-۶. نمونه گیری
۲۷.....	۳-۷. مشکلات زبان فارسی در وب کاوی
۲۸.....	۳-۷-۱. خصوصیات وب های فارسی از نظر زبان
۲۸.....	۳-۷-۲. روش هایی برای بهبود کاوش وب های فارسی
۲۹.....	۳-۸. خلاصه فصل
۳۰.....	فصل چهارم - پیش بینی سلامت با وب
۳۰.....	۴-۱. مقدمه
۳۲.....	۴-۲. نظارت اپیدمی
۳۳.....	۴-۳. تجزیه و تحلیل نفوذ اجتماعی
۳۵.....	۴-۴. روند یابی
۳۶.....	۴-۵. تجزیه و تحلیل پیش بینی شده اجتماعی (TPA)
۳۸.....	۴-۶. مدل موضوعی بیماری (ATAM)
۳۹.....	۴-۵. خلاصه فصل
۴۰.....	فصل پنجم - نتیجه گیری و جمع بندی
۴۰.....	۵-۱. مقدمه

۴۱	۲-۵. مروری بر فصل های گذشته
۴۲	۳-۵. مروری بر تحلیل و نقد ادبیات موضوع
۴۲	۴-۵. محدودیت ها و نکات قابل توجه
۴۳	۵-۵. زمینه ها و موضوعات پژوهشی برگرفته از مطالعه
۴۳	۶-۵. طرح و بیان مسئله تحقیق پایان نامه
۴۳	۷-۵. زمینه های کاربردی پیشنهادی و میزان دسترسی به داده های تحقیق
۴۳	۱-۷-۵. استراتژی تحقیق
۴۴	۲-۷-۵. روش گردآوری داده ها و جامعه آماری
۴۴	۸-۵. خلاصه فصل
۴۵	فهرست منابع
۵۳	پیوست ۱- فهرست اصطلاحات فارسی به انگلیسی
۵۴	پیوست ۲- فهرست اصطلاحات انگلیسی به فارسی

فصل اول - معرفی و کلیات

۱-۱. مقدمه

نسل اول اینترنت شامل صفحات استاتیکی بود که به کاربران ارائه می شد. همانطور که تکنولوژی اینترنت تکامل یافت، برنامه های جدید نیز اجازه تعامل بیشتر، شخصی سازی و همکاری بین کاربران ایجاد کرد. اصطلاح Web2.0 اولین بار توسط Tim O'Reilly در سال ۲۰۰۵ برای اشاره به این نسل جدید از فن آوری های وب، ابداع شد (Meier و همکاران، ۲۰۰۷). این فن آوری ها شامل سایت های شبکه های اجتماعی که مشوق همکاری و بحث و گفتگو میان کاربران است، سایتهایی که توسط دانش جمعی ایجاد شده است مانند ویکی پدیا، و mashup ها، که اجازه می دهند برنامه های کاربردی وب به راحتی برای ایجاد ارزش بیشتر ترکیب شوند، می شوند (Chun و همکاران، ۲۰۱۲). برنامه های کاربردی همراه، با افزایش در دسترس بودن دستگاه های تلفن همراه که اینترنت دارند (مانند نوت بوک و آی فون اپل)، باعث شده است که اینترنت در حال حاضر واقعا بخشی جدایی ناپذیر از زندگی روزمره باشد. وب اجتماعی رسانه ای خصوصا پویا است که پالس های جامعه را در زمان واقعی سنجش میکند. وبلاگ ها، میکرو وبلاگ مانند توییتر (<http://www.twitter.com>)، و سایت های شبکه های اجتماعی مانند فیس بوک (<http://www.facebook.com>)، مردم را قادر به انتشار داستان های شخصی خود، نظرات، نظرات در مورد یک محصول و موارد دیگر در زمان حال می کند. آنها همچنین می توانند برچسب جغرافیایی و گزارش در مورد مکان فعلی خود را به اشتراک بگذارند و اجازه می دهند دیگران به دنبال مکان آنها، همه در زمان حال، با استفاده از خدمات اجتماعی محل آگاه مانند Microsoft Vine (<http://www.vine.net>) باشند. چنین برنامه های کاربردی دارای قدرت بیشتری می شوند وقتی که از طریق mashup تجمیع می شوند. چنین ابزارهایی داده ها را از منابع نامتجانس در یک منبع داده ترکیب می کنند (Kamel Boulos و دیگران، ۲۰۰۸؛ Wheeler و دیگران، ۲۰۰۷). یک مثال مفید از mashup، HealthMap است که نقشه جهانی هشدار بیماری است (<http://www.healthmap.org>). برنامه هایی از web2.0 که در زمینه مراقبت های بهداشتی است فن آوری Health2.0 یا Medicine2.0 نامیده می شوند (Hughes، ۲۰۰۸؛ Eysenbach، ۲۰۰۸). برنامه های کاربردی

Health2.0 اجازه می دهد بیماران در جمع آوری و ایجاد اطلاعات همکاری کنند. اثر مثبت Health2.0

افزایش دانش پزشکی در اینترنت است، که بیشتر آن توسط خود بیماران می باشد. با گسترش وبلاگ ها، انجمن های بحث و گفتگو، اخبار و خدمات آنلاین بخشی از محتوای وب سایت ها که مربوط به بهداشت و پزشکی است نیز در حال گسترش است. و نشان داده شده که محتوای پزشکی به روز وب سایت ها یک منبع اطلاعات با ارزش و ماندگار برای فرضیه های پزشکی و تست آنها می باشند (حتی با وجود خطا دار بودن آنها) (MacLean و Seltzer، ۲۰۱۱).

با توجه به این امر، اینترنت اساسا راهی که بیماران برای به دست آوردن اطلاعات مربوط به مراقبت های بهداشتی و پزشکی داشتند تغییر داده است. بیماران، که تنها تکیه بر متخصصان مراقبت های بهداشتی برای مشاوره و درمان داشتند، در حال حاضر به دنبال اطلاعات بهداشت و درمان خود از اینترنت هستند (Chun و همکاران، ۲۰۱۲). برنامه های کاربردی وب که می توانند این اطلاعات پراکنده را ادغام و به کاربر ارائه کنند، نقش مهمی را بر عهده دارند.

۲-۱. بیان مسئله و اهداف مطالعه

همه گیری و همه جایی بودن اینترنت و دسترسی افراد به بسیاری از منابع اطلاعاتی، یک روش غیر رایج نظارت در بهداشت و سلامت است (Corley و همکاران، ۲۰۱۰).

در این سمینار به بررسی چگونگی اتصال داده های پراکنده مرتبط با سلامت از جوامع مختلف وب و ارائه دانش یکپارچه اطلاعات سلامت (به طور خاص از داده های رسانه های اجتماعی مبتنی بر جوامع بیماران، سایت های مراقبت درمانی با محتوای تخصصی)، به منظور نظارت بر بهداشت عمومی و پیش بینی اتفاقاتی مثل شیوع بیماری ها از طریق روش های همچون داده کاوی متن و ساختار وب و رسانه های اجتماعی (WSM) (Corley و همکاران، ۲۰۱۰) که می تواند یک منبع نظارتی برای بیماری فراهم می کند و می تواند جوامع آنلاین برای ارتباطات هدفمند بهداشت عمومی (PHC) را شناسایی کند، سیستم بازیابی اطلاعات وب (WebIRS)، که به متخصصان مراقبت های بهداشتی برای به دست آوردن اطلاعات به روز دانش پزشکی و اطلاعات از طریق شبکه جهانی وب (WWW) کمک می کند (Ting و همکاران، ۲۰۱۳) و تجزیه و تحلیل پیش بینی اجتماعی (TPA) (Kamel Boulos و همکاران، ۲۰۱۰)، پرداخته می شود.

۳-۱. اهمیت موضوع

متن کاوی جستجوها در گوگل نشان داده است که مطالب ارسالی در مورد آنفلونزا با روند بیماری آنفلونزا در دنیای واقعی یکسان است (Corley و همکاران، ۲۰۱۰) و مطالعات متعددی با استفاده از تویتر برای

رديابی روندهای مختلف انجام شده اند (Paul و Dredze، ۲۰۱۲). همچنین در رابطه با سلامت به طور خاص، سیستم های ارتباطی در رسانه های اجتماعی و وب برای به دست آوردن اطلاعات مفید بهداشتی و سلامت عمومی خصوصا با توجه به اینکه کاربران نهایی می توانند از طریق دستگاه های تلفن همراه و وسایل مشابه به اینترنت به طور مداوم متصل شوند (Steele، ۲۰۱۱) و اطلاعاتی از سلامت و بهداشت خود منتشر کنند، نقش مهمی دارند. همچنین سکوهایی مانند تویتر برای محققان فرصت کافی برای تحلیل پدیده های اجتماعی فراهم آورده است. استفاده از این ظرفیت جهت نظارت عمومی بر بهداشت و سلامت با توجه به اینکه در ایران نفوذ اینترنت در حال افزایش است، اهمیت بالایی پیدا کرده است. البته چالش هایی در این زمینه وجود دارد از جمله اعتبار سنجی اطلاعات و همچنین مشکلاتی که در رابطه با زبان فارسی در داده کاوی وب وجود دارد. زبان فارسی، در تلاقی با جهان الکترونیکی، بخصوص از بعد رسم الخط، دارای مشکلاتی است که کاوش در محتویات آن را دچار کم کیفیتی می نماید (شهیدی، ۱۳۸۴).

۱-۴. اهداف مطالعه

اهداف مطالعه شامل موارد زیر است:

۱. راه اندازی سیستم جمع آوری اطلاعات وبلاگ های فارسی زبان
۲. استفاده از روندیابی گوگل و api گوگل جهت کسب اطلاعات نظارتی سلامت
۳. راه اندازی سیستم نقشه سلامت ایران بر مبنای اطلاعات کسب شده از سیستم های خبری
۴. ایجاد پایگاه کلمات کلیدی در حوزه سلامت وب به زبان فارسی

۱-۵. تاریخچه موضوع

تحقیقات مختلفی در زمینه استفاده از منابع مختلف وب و روش های مختلف استخراج اطلاعات از وب جهت شناسایی بیماری ها و شیوع بیماری ها انجام شده است. از مهم ترین این منابع وب سایت تویتر (<http://www.twitter.com>) است که در رابطه با آنفلونزا و شیوع آن تحقیقات مختلفی صورت گرفته است (Paul و Dredze، ۲۰۱۲) و در تحقیق دیگری رسانه های اجتماعی به عنوان یک ابزار جدید برای اندازه گیری سلامت روانی و نظارت معرفی شده اند (Choudhury و همکاران، ۲۰۱۳). همچنین در مورد شیوع وبا در هائیتی از طریق وب سایت HealthMap (<http://www.healthmap.org>) و تویتر تحقیقاتی انجام شده است (Chunara و همکاران، ۲۰۱۲) که نشان دهنده همبستگی مناسب اطلاعات بدست آمده با اطلاعات گرفته شده از منابع رسمی مانند MSPP که یک نهاد دولتی است دارد. در فرایند استخراج اطلاعات، به ابزار مناسبی جهت نمونه گیری و نحوه نمونه گیری از داده های منابع مختلف وب

خصوصاً توئیتر پرداخته شده (Choudhury و همکاران، ۲۰۱۰) که نشان دهنده اهمیت این منبع برای استفاده در پیش بینی ها بر اساس محیط اجتماعی وب است.

در روش های تحلیل داده های مختلف وب، روش تجزیه و تحلیل پیش بینی اجتماعی (TPA) بررسی شده است (Kamel Boulos و همکاران، ۲۰۱۰). این روش بر روی "داده کاوی" در پست ها و پرسش های کاربران در محیط SocialWeb که شامل ابزارهایی مانند وبلاگ، شبکه های اجتماعی، میکروبلاگ، پرس و جو، تالار گفتمان، چت و موارد مشابه اجتماعی دیگر است، متمرکز شده است.

در تحقیقات اخیر (Kizito و دیگران، ۲۰۱۳) نشان داده شده که سیستم الکترونیکی نظارت بر بیماری ها منجر به بهنگام بودن و دقت بالاتر روش های شناسایی شیوع بیماری است و در تحقیق دیگری (Ochneva، ۲۰۱۳) استفاده از تکنولوژی مکان محور بر اساس محیط وب برای شناسایی آلرژی های مختلف بررسی شده است. همچنین King و دیگران (۲۰۱۳) به تاثیر توئیتر در سیستم ملی سلامت در انگلستان پرداخته اند.

۱-۶. تعاریف و اصطلاحات

بلاگ (Blog):

سایتی شامل یک مجله بر خط شخصی که در آن نظرات و رویدادها منتشر می شود. مختصر شده کلمه وبلاگ. استفاده شده از سال ۱۹۹۹ (دیکشنری Merriam-webster). اطلاعاتی منظم شامل افکار، نظرات یا تجربیات که بر روی اینترنت برای خواندن قرار می گیرد (دیکشنری Cambridge).

بیماری (Disease):

شرایطی برای موجودات زنده در بدن یا یکی از اعضای آنها که باعث اختلال در عملکرد طبیعی می شود و معمولاً علائم و نشانه هایی دارد (دیکشنری Merriam-webster). مریضی افراد، حیوانات، گیاهان و غیره که ناشی از عفونت یا خطای سلامتی به واسطه تصادف باشد (دیکشنری Cambridge).

داده کاوی (Data Mining):

عملیاتی جهت جستجو در حجم زیادی از اطلاعات رایانه ای برای پیدا کردن روندها و الگوها. استفاده شده از سال ۱۹۸۸ (دیکشنری Merriam-webster). پردازش حجم زیادی از اطلاعات ذخیره شده در رایانه برای دیدن الگوها، تغییرات و موارد دیگر (دیکشنری Cambridge).

شبکه اجتماعی (Social Network):

سایت یا برنامه ای رایانه ای که به مردم اجازه ارتباط یا به اشتراک گذاری اطلاعات در اینترنت را توسط رایانه یا تلفن همراه می دهد (دیکشنری Cambridge).

شناسایی (Detect):

عملی برای کشف یا تعیین وجود یک واقعیت (دیکشنری Merriam-webster). عمل متوجه شدن یا کشف چیزی (دیکشنری Cambridge).

وب (Web):

مختصر شده عبارت world wide web به معنای شبکه گسترده جهانی (دیکشنری Merriam-webster).

ویکی (Wiki):

سایتی که به بازدیدکنندگان اجازه تغییر، توسعه و تصحیح اطلاعات را می دهد. استفاده شده از سال ۱۹۹۵ (دیکشنری Merriam-webster). سایتی که به کاربران اجازه اضافه، حذف و ویرایش محتوا را می دهد (دیکشنری Cambridge).

۷-۱. علائم و اختصارات

ATAM: Ailment Topic Aspect Model
CDC: Center for Disease Control and Prevention
IE: Information Extractor
ILI: Influenza Like Illness
KEF: Knowledge Encapsulation Framework
LDA: latent Dirichlet allocation
MedISys: Medical Information System
PHC: Public Health Communications
PULS: Pattern-based Understanding and Learning System
RDF: Resource Description Framework
RSS: Rich Site Summary
UMLS: Unified Medical Language System
WebIRS: Web Information Retrieval System
Wiki: What I Know Is
WSM: Web and Social Media
WWW: World Wide Web

۸-۱. ساختار کلی گزارش

در فصل اول راجع به تحولات وب و اهمیت موضوع شناسایی بیماری از طریق وب اشاره شده و در ادامه در فصل دوم منابع مختلف اطلاعات سلامت بر روی وب بررسی می شوند که این منابع شامل: شبکه های اجتماعی، وبلاگ ها، اخبار، روند های جستجو و میکرو بلاگ ها می باشند.

در فصل سوم به چالش های استخراج اطلاعات از این منابع و مشکلات زبان فارسی پرداخته شده است و به انواع سیستم های خزنده (Crawler) اشاره می شود.

در فصل چهارم به روند های پیش بینی بر اساس اطلاعات استخراج شده از منابع مختلف وب پرداخته می شود و روش های همچون داده کاوی متن و ساختار وب و رسانه های اجتماعی (WSM)، سیستم بازیابی اطلاعات وب (WebIRS) و تجزیه و تحلیل پیش بینی اجتماعی (TPA) بیان می شود.

در فصل آخر جمع بندی فصول گذشته و زمینه های کاربردی این تحقیق بیان شده است.

۹-۱. خلاصه فصل

با شروع انقلاب Web2.0، حجم عظیمی از اطلاعات مختلف از جمله اطلاعات سلامت و بهداشت کاربران بر روی وب سایت های مختلف منتشر شد. این اطلاعات می تواند به عنوان مبنایی برای نظارت بر شیوع بیماری های مختلف قرار گیرد. در ادبیات موضوع، موارد مختلفی در این زمینه اشاره شده است که اکثرا در آمریکا و بعضی در اروپا و کشورهای دیگر انجام شده است.

اما چنین تحقیقاتی در ایران و با توجه به مشکلات زبان فارسی در داده کاوی، هنوز انجام نشده است که می تواند زمینه ی مناسبی برای تحقیق باشد.

در این فصل ابتدا مقدماتی در مورد Web2.0 بیان شد و در ادامه به بیان مساله و اهمیت موضوع پرداخته شده و در نهایت با بیان خلاصه ای از ادبیات موضوع و تعاریف و اصطلاحات فصل به اتمام رسید.

فصل دوم - منابع داده های سلامت در وب

۲-۱. مقدمه

شبکه جهانی وب (WWW) یکی از منابع مفید و رایج برای متخصصان پزشکی برای به دست آوردن و به روز آوری دانش است (Bell و Sethi، ۲۰۰۱). در سال های اخیر وب به عنوان یک پلت فرم اجتماعی که در آن مصرف کنندگان سلامت به طور جدی به ایجاد و مصرف محتوای آن روی آورده اند، تبدیل شده است (Luque و همکاران، ۲۰۱۱). با بروز فناوری Web2.0 سایت ها و برنامه های جدیدی وارد عرصه وب شد که شامل سایت های شبکه های اجتماعی مانند facebook، سایتهایی که توسط دانش جمعی اداره می شود مانند ویکی پدیا، وبلاگ هایی که اطلاعات مختلفی از کاربران منتشر می کند مانند Blogger و میکرو وبلاگ هایی مانند Twitter که معمولاً عباراتی کوتاهی را از افراد را منتشر می کنند که معمولاً نشان دهنده وضعیت و افکار نویسنده است.

برنامه هایی از web2.0 که در زمینه مراقبت های بهداشتی هستند، فن آوری Health2.0 یا Medicine2.0 نامیده می شوند که اجازه می دهند بیماران در جمع آوری و ایجاد اطلاعات همکاری کنند. اثر مثبت Health2.0 افزایش دانش پزشکی در اینترنت است، که بیشتر آن توسط خود بیماران می باشد که با گسترش وبلاگ ها، انجمن های بحث و گفتگو، اخبار و خدمات آنلاین بخشی از محتوای وب سایت ها که مربوط به بهداشت و پزشکی است نیز در حال گسترش است. این اطلاعات، به عنوان "محتوای پزشکی عمومی" (PML) شناخته می شوند. به عنوان مثال وبلاگ اختصاص داده شده به تجربه بیماری با درمان سرطان تجربی، انجمن های بحث آنلاین برای مادران باردار و مقاله ای شامل اخبار مربوط

به Asprin انواع مختلف PML هستند. اگر چه این اطلاعات ممکن است خطا داشته باشند اما PML ذاتا ارزشمند است.

این امر با توجه به اینکه نیاز به روز رسانی اطلاعات پزشکی و دانش به ویژه در بهداشت تخصصی حیاتی است (Rogers و همکاران، ۲۰۱۲)، اهمیت منابع اطلاعات سلامت در وب را دو چندان می کند. ماهیت پویا و ویژگی به روز رسانی مداوم وب اجتماعی آن را به محیط مناسبی برای جمع آوری اطلاعات در رشته های مختلف تبدیل می کند و کاربران را قادر به استفاده از "حکمت استفاده از جمعیت" می کند (Kamel Boulos و همکاران، ۲۰۱۰). علاوه بر این، وب اجتماعی برای خوشه چینی از توده جمعی در مورد مسائل فعلی، رویدادها و محصولات مفید است (Kamel Boulos و دیگران، ۲۰۰۷). تحقیقات اخیر به اثر استفاده از وب برای گرد هم آوردن جوامع از بیماران، خانواده ها و پزشکان در سایت های مانند CureTogether و برای تشخیص شیوع بیماری ها، همانطور که در Google Flu و Biosurveillance نشان داده شده ، اشاره شده است (MacLean & Seltzer، ۲۰۱۱). در این فصل به بررسی منابع مختلف اطلاعات سلامت در وب پرداخته می شود.

۲-۲. شبکه های اجتماعی (Social Network)

یکی از منابع بهینه و مناسب جهت کسب اطلاعات سلامت در وب اطلاعات شبکه های اجتماعی مانند facebook می باشد. این سایت ها اجازه جمع اطلاعات و تجارب بیماران را می دهد که ساخت یافته تر از تالار های گفتمان است. همانطور که اشاره شد در PatientsLikeMe اطلاعات بیماران جمع آوری شده و به عنوان گزارش های بصری برای شرکت کنندگان ارائه می گردد (Wicks و همکاران، ۲۰۱۰)، که باعث ایجاد یک مخزن با ارزش از اطلاعات پزشکی شده است.

شبکه های آنلاین اجتماعی و رسانه های اجتماعی قابلیت های مهمی در استفاده در مراقبت های مرتبط با سلامت دارند. این پتانسیل در افزایش شیوه های ارتباط سلامت مطرح می شود. این پتانسیل که به نفع بهداشت عمومی و کمک به پیشگیری از برخی از بیماری های مزمن است، با توجه به اهمیت انتخاب شیوه زندگی برای شرایط مزمن استفاده می شود، اما تحقیقات قابل توجهی هنوز هم مورد نیاز است (Steele، ۲۰۱۱).

این نوع ارتباطات از دو جنبه ویژگی ها، نوع تعاملات و حالت های ارتباطی متمایز می شوند.

۲-۲-۱. ویژگی‌های مربوط به رسانه‌های اجتماعی

• ارتباط یک به چند: بر خلاف تعدادی از نتایج انقلاب ارتباطات الکترونیکی در دهه‌های اخیر، از جمله ایمیل، پیام‌های فوری، پیام‌های متنی و موارد دیگر، رسانه‌های اجتماعی بر روی مدل یک به یک متمرکز نشده است. این به این معنی است که ارتباطات بیشتر بر روی گروه‌های ارتباطی متمرکز شده است (Boyd و همکارانش، ۲۰۱۰؛ Java و همکاران، ۲۰۱۰؛ Perreault و Ruths، ۲۰۱۱).

• ماهیت اجتماعی: به دلیل مشخصه‌هایی که ذکر شد، و تا حدودی به دلیل این واقعیت که ارتباطات بین کسانی که در اجتماع قرار می‌گیرند و یا متخصص هستند و یا سایر روابط در دنیای واقعی و یا شخصی دارند، ایجاد می‌گردد؛ ارتباط یکی از جنبه‌های بسیار قوی و جذاب است (Kumar).

• مانع کم در ارتباطات: در حالی که اشکال سنتی ارتباط مانند نامه نیاز به هر دو فاکتور عملیاتی و اجتماعی دارد که موانع بیشتر مربوط به برقراری ارتباط به محل به صورت فیزیکی است، ارتباطات الکترونیکی و اخیراً رسانه‌های اجتماعی و شبکه‌های اجتماعی منجر به کم شدن موانع در ارتباطات شده‌اند (Skeels و Grudin، ۲۰۰۹). این امر به ویژه با ترکیب این فن‌آوری ارتباطی با فن‌آوری‌های دستگاه تلفن همراه از جمله گوشی‌های هوشمند باعث گسترش فزاینده‌ی این ارتباطات شده است (Perreault و Ruths، ۲۰۰۱).

• همبستگی جغرافیایی: در حالی که رسانه‌های اجتماعی و شبکه‌های اجتماعی از ارتباطات مستقل از مکان حمایت می‌کنند، این شبکه‌ها به عنوان قابلیت اتصال شبکه‌های اجتماعی بین مردم شناخته شده‌اند که این ارتباطات به طور معمول نسبت بالاتری از مخاطبین جامعه خود و یا شهر خود را شامل می‌شود (Onnela و همکاران، ۲۰۱۰).

۲-۲-۲. طبقه‌بندی تعاملات اجتماعی شبکه بهداشت

این دسته بندی شامل پنج دسته عمده بر اساس گروه‌های حاضر در تعامل است:

- تعاملات بیمار و بیمار
 - تعاملات بیمار و درمانگاه
 - تعاملات سلامت و درمانگاه عمومی و مصرف‌کننده
 - تعاملات پژوهشگر و بیمار
 - تعاملات شرکت‌ها و بیمار
- در شبکه‌های اجتماعی گروه‌ها افراد هستند. اگر چه برخی از گروه‌های فردی به نمایندگی از سازمان‌ها در سیستم وجود دارند.

۲-۳. حالت های ارتباطاتی با رسانه های اجتماعی

- شیوه های رسانه های اجتماعی و ارتباطات شبکه های اجتماعی عوامل موثری در ارتباطات سلامت دارد:
- مشاهده متمرکز روزآمد: با توجه به جنبه های عمومی اشکال مختلف ارتباطات رسانه های اجتماعی، این ارتباطات یک منبع ارزشمند از اطلاعات کل جامعه است.
 - ارتباط یک به چند: این اثر شامل رفتار ارتباطی بیمار و شیوه های تعامل بیماران و مصرف کنندگان است.
 - ارتباط یک به یک: ارتباط سلامت بسیار صمیمی، فوری و شخصی فعال شده بین بیماران
- این نوع تعاملات اجتماعی با فعال کردن محیط فناوری برای بهبود شبکه های اجتماعی سلامت، افراد می توانند با شبکه های اجتماعی آنلاین از طریق هر دستگاه رایانه ی متصل به اینترنت ارتباط برقرار کنند. با این حال، برنامه های کاربردی رسانه های اجتماعی بهداشت معمولاً از نزدیک به فرد فیزیکی مرتبط است و دارای قابلیت های بیشتری از طریق این اتصال فیزیکی است. به این ترتیب، در حالی که ما هر یک از مقوله های فوق را در نظر می گیریم، دو توانمندساز فن آوری زیر باعث افزایش این تعاملات شده اند:
- دستگاه های تلفن همراه: که باعث ارتباط مستمر با شبکه های اجتماعی شده است.
 - سنسور: اتصال به فیزیولوژی جسمی شخص که اطلاعات مختلفی را به سیستم های اجتماعی منتقل می کند (KO و همکاران، ۲۰۱۰).

۲-۳. میکرو بلاگ (Microblog)

از منابع مهم جهت کسب اطلاعات در خصوص سلامت افراد میکرو بلاگ هایی مانند twitter می باشد. در این بلاگ های وضعیت های مختلف کاربران به صورت روزانه از طریق محتوایی که منتشر می کنند، قابل سنجش است. این نوع ارتباطات با توجه به افزایش نسبت ارتباطات میکرو وبلاگ نویسی مانند کسانی که از طریق توییترو توسط دستگاه های تلفن همراه ارتباطات برقرار می کنند، به سرعت در حال افزایش است (Parr، ۲۰۱۱).

چنین ارتباطی می تواند در راستای گسترش سلامت عمومی قرار گیرد و طیف وسیعی از برنامه های کاربردی بهداشت عمومی در توییترو ایجاد گردد (Paul و Dredze، ۲۰۱۱).

از طریق چنین محیط هایی می توان اعمال مختلفی را مانند ردیابی بیماری در طول زمان (نظارت سندرومی)، اندازه گیری عوامل خطر رفتاری، بومی سازی بیماری بر اساس منطقه جغرافیایی، و تجزیه و تحلیل علائم و استفاده از دارو، انجام داد که در تحقیقات مختلف نشان داده شده که توییترو دارای کاربرد گسترده ای برای تحقیقات بهداشت عمومی است.

توییتر، فیس بوک و دیگر رسانه های اجتماعی کاربران خود را تشویق می کنند تا افکار، عقاید و جزئیات زندگی خود را بیان کنند. توییت ها و به روز رسانی وضعیت کاربر می تواند جزو رویدادهای مهم یا به نظر پوچ باشد. بسیاری از پیام ها دارای ارزش اطلاعاتی کمی هستند اما تجمیع میلیون ها پیام می تواند دانش مهمی تولید کند.

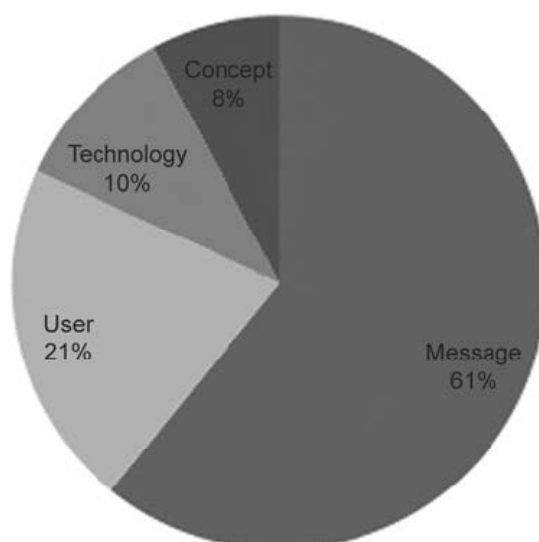
به طور مشابه، توییت ها به موضوعات مرتبط با سلامت، مانند "من تب کرده ام، من آنفلوآنزا گرفته ام" و موارد مشابه اشاره کرده است. این توییت ها نشان می دهد که کاربر دارای بیماری (آنفلوآنزا) علائم مرتبط (تب، و غیره) و درمان (تیلنول) است. خود گزارش دهی بهداشتی میلیون ها نفر از کاربران می تواند اطلاعات گسترده ی مربوط به سلامت جامعه به صورت روزانه فراهم کند (Paul و Dredze، ۲۰۱۲).

در تحقیق دیگری پتانسیل رسانه های اجتماعی به عنوان یک ابزار جدید برای اندازه گیری سلامت روانی و نظارت پرداخته شده است. محیط هایی مانند توییتر و فیس بوک به طور فزاینده ای در حال گسترش است و افراد در آن به اشتراک گذاشتن افکار و احساسات خود در سراسر اتفاقات در زندگی روزمره می پردازند. در این تحقیق نشان داده شده که احساسات و زبان مورد استفاده در رسانه های اجتماعی ممکن است احساس بی ارزشی، احساس گناه، درماندگی، و خود نفرت که مشخص کننده افسردگی در زندگی است (Choudhury و همکاران، ۲۰۱۳) نشان دهد.

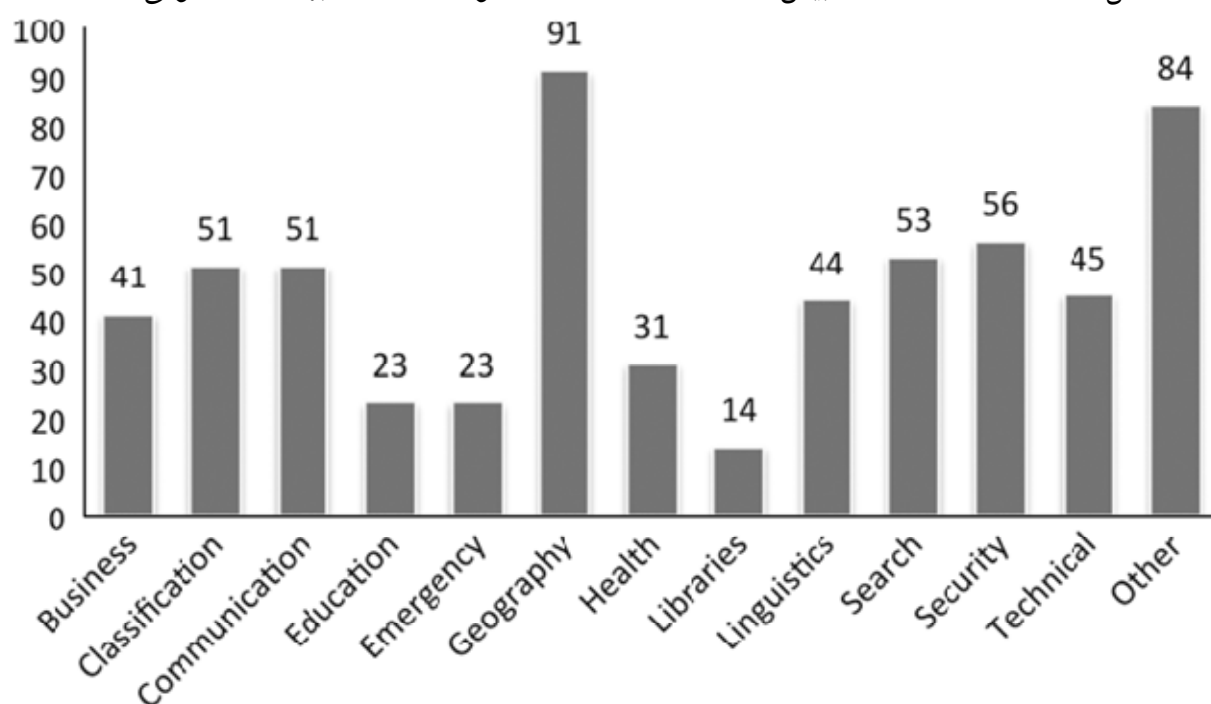
تکیه بر رسانه های اجتماعی به عنوان یک ابزار ارزیابی سلامت رفتاری دارای مزیت های دیگری نیز هست. به عنوان مثال، در مقابل روش خود گزارش در بررسی های رفتاری، که در آن پاسخ ها توسط آزمونگر تعیین می شود و به طور معمول شامل عباراتی و حقایقی (گاهی اوقات ذهنی) از سلامت است، اندازه گیری از طریق رسانه های اجتماعی از رفتار، فعالیت های اجتماعی و نحوه بیان در یک محیط طبیعی سنجش می شود. چنین فعالیت هایی در زمان واقعی است، و در دوره ای از زندگی روزمره شخص اتفاق می افتد. از این رو این ارزیابی به خاطر اثر حافظه یا نظر آزمایشگر کمتر آسیب پذیر است.

در جمع آوری اطلاعات از محیط های میکرو بلاگ چالش اولیه جمع آوری اطلاعات کافی از رسانه های اجتماعی است که یک نشانگر مناسب برای سنجش سلامت فرد و جامعه باشد.

در تحقیقات اخیر (Williams و دیگران، ۲۰۱۳) نشان داده شده که از سال ۲۰۰۶ به بعد بیش از ۲۳۰ هزار مقاله و یا مطلب علمی (به نقل از Google Scholar) در مورد میکرو بلاگ و خصوصا توییتر منتشر شده است که همانطور که در شکل ۱-۲ آمده است بیش از نیمی از تحقیقات حول کشف اطلاعات از مطالب منتشر شده در توییتر می باشد که نشان دهنده اهمیت این منبع اطلاعاتی در کشف دانش و همچنین نظارت است.



تصویر ۱-۲: نسبت محتوای بررسی شده در مقالات منتشر شده با استفاده از توییتر (Williams و همکاران، ۲۰۱۳) و در شکل ۲-۲ نشان داده شده که بیش از ۵ درصد مقالات در حوزه سلامت و بهداشت عمومی است.



تصویر ۲-۲: نسبت موضوع مقالات منتشر شده با استفاده از توییتر (Williams و همکاران، ۲۰۱۳)

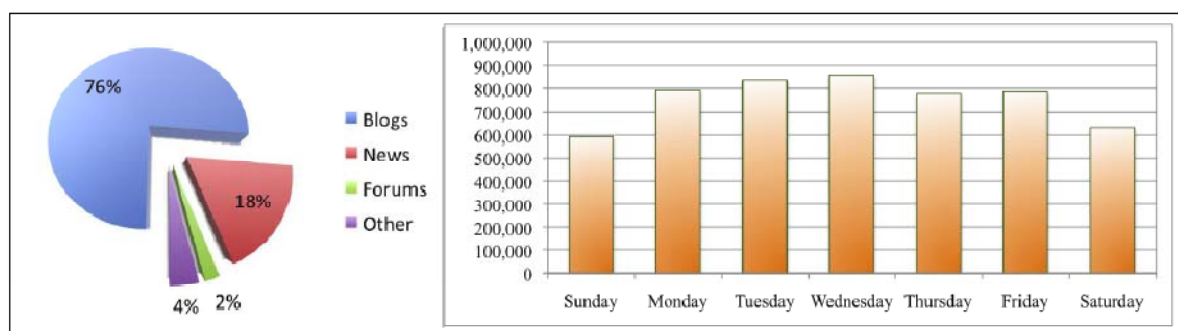
۴-۲. وبلاگ ها و اخبار (News & Blogs)

یکی دیگر از منابع جهت کسب اطلاعات سلامت عمومی از طریق وب، محتوای وبلاگ ها و اخبار منتشر شده در سایت های خبری است. البته این اطلاعات دارای حجم زیادی هستند و پردازش این اطلاعات نیاز به روش هایی برای بهبود داده کاوی در این اطلاعات دارد. این سایت ها در نقطه میانی بین سایت های مراقبتی و تالار های گفتمان قرار گرفته است. وبلاگ ها توسط هر فردی نوشته شده است، از بیماران یا

متخصصان پزشکی. آنها اهداف مختلفی دارند از جمله پژوهش، چشم انداز بهداشت و درمان و حتی انتشار نسخه یک بیماری (Chun و همکاران، ۲۰۱۲).

وب سایت و رسانه های اجتماعی (WSM) ارائه یک منبع برای تشخیص افزایش در ILI (بیماری های همچون آنفلوآنزا که امکان شیوع دارند) است. در مقالات متعددی نشان داده شده که مطالب وبلاگ ها نیز به عنوان یکی از منابع شناسایی روند آنفلوآنزا قابل استفاده است (Corley و همکاران، ۲۰۱۰).

برای مثال شخص می تواند یکی از نشانه های بیماری خاص را جستجو کند. اغلب نتایج حاصل از چنین پرس و جویی به سایت هایی است که ممکن است کمک به تشخیص بیماری و ارائه توصیه های پزشکی باشد (به عنوان مثال، PeopleLikeMe.com، WebMD.com). شکل ۲-۳ نسبت حجم اطلاعات ارسالی در وب در محیط های مختلف اجتماعی را نشان می دهد که نشان می دهد حجم بسیاری از این اطلاعات در وبلاگ ها منتشر می شود.



تصویر ۲-۳: اطلاعات منتشر شده در وب اجتماعی به تفکیک نوع منبع و روز های انتشار از ۵ اکتبر الی ۲۱ مارس ۲۰۰۹ (Corley و همکاران، ۲۰۱۰)

برای مثال HealthMap که یک نقشه سلامت عمومی است، محیطی برای نظارت خودکار است که به طور مداوم در حال شناسایی مشخصات و نقشه حوادث بهداشت عمومی و پزشکی است، از جمله شیوع بیماری و بیماری های همه گیر. ۱۲ منبع اطلاعات برای HealthMap شامل منابع رسانه های خبری و گروه های بحث است (Chunara و همکاران، ۲۰۱۲).

۲-۵. تالارهای گفتار (Forums)

با ظهور اینترنت، اساسا راهی که بیماران برای به دست آوردن اطلاعات مربوط به مراقبت های بهداشتی و پزشکی داشتند تغییر کرد و بیماران، که تنها تکیه بر متخصصان مراقبت های بهداشتی برای مشاوره و درمان داشتند، در حال حاضر به دنبال اطلاعات بهداشت و درمان خود از اینترنت هستند (Chun و همکاران، ۲۰۱۲). اطلاعات پزشکی و سلامتی در اینترنت توسط بسیاری از منابع ارائه شده است: وب سایت بیمارستان، وب سایت رسمی انجمن پزشکی، پورتال بهداشت و درمان، سایت های شبکه های اجتماعی،

وبلاگ ها نوشته شده توسط بیماران و متخصصان پزشکی و تالارهای بحث آنلاین و تالارهای گفتمان. بر اساس یک نظرسنجی که در ۲۰۱۱ توسط Pew Internet and American Life Project انجام شده است، ۵۹ درصد از تمام بزرگسالان آمریکایی در اینترنت به جستجو برای اطلاعات مرتبط با سلامت هستند (Fox, ۲۰۱۱).

یکی از محیط های مناسب برای اطلاعات سلامت تالار های گفتمان است. تالارهای گفتمان فضایی است که در آن کاربران پرسش و پاسخ های مختلفی را انجام می دهند که معمولاً شامل وضعیت، نیاز و خواسته کاربر است. از طریق محتوای این نوع سایت ها می توان به سنجش دغدغه ها و در نهایت وضعیت سلامت کاربران پرداخت و در صورتی که تالار های بررسی شده دارای تعداد کافی از کاربران باشند می توانند به عنوان ابزار مناسبی جهت نظارت بر سلامت از آنها استفاده کرد.

در تحقیقاتی بیان شده که سیستم سنجش سلامت از اطلاعات PatientsLikeMe، WebMD و PubMed به صورت یکپارچه استفاده می کند (Chun و همکاران، ۲۰۱۲). سایت های تالار گفتمان ویژگی هایی دارند که باعث طبقه بندی نوع آنها و تمایز آنها با دیگر منابع سلامت در وب است.

۲-۵-۱. ویژگی های تالار های گفتمان بهداشت

توجه به انواع مختلف وب سایت های سلامت که هر یک برای پیدا کردن اطلاعات مربوطه چالش هایی دارند، مهم است. در ادامه شرحی از انواع تالارهای گفتمان سلامت بررسی می گردد:

- سایت های مراقبتی: این سایت ها که توسط برخی از نهاد رسمی حمایت می شوند، اغلب توسط یک سازمان دولتی، سازمان های مربوط به بیماری، یا سازمان پزشکی راه اندازی می شوند. به عنوان مثال، PubMed.gov توسط کتابخانه ملی پزشکی ایالات متحده ایجاد شده است. این سایت در بیش از ۲۱ میلیون استناد مقالات پژوهشی منتشر شده در پزشکی اشاره شده است. این سایت جستجو بر اساس کلید واژه های سرعنوان های موضوعات پزشکی (MeSH) را فراهم می کند.

- سایت های پرسش پاسخ مدیریت شده: در این ساختار، بیماران سوالاتی در مورد وضعیتشان می پرسند و متخصصان پزشکی پاسخ می دهند. پرسش و پاسخ ها به صورت متن های بدون ساختار خاص مانند تالار های گفتمان پزشکی است. با این حال، از آنجا که اطلاعات به صورت پرسشنامه است قالب محدود تر است. یک مثال از چنین سایتهایی، سایت انجمن سرطان پستان WebMD.com است که توسط یک متخصص سرطان پستان مدیریت می شده است.

- تالار های گفتمان آزاد: در سایت های مراقبتی و پرسش و پاسخ، انتقال اطلاعات توسط کارشناسان به غیر کارشناسان است. در انتهای دیگر این طیف، تالار های گفتمان وجود دارد که اجازه می دهد بیماران با

هم ارتباط برقرار کنند و به یکدیگر ارائه اطلاعات کنند. نمونه های بسیاری از تالار های گفتمان مرتبط با سلامت در اینترنت وجود دارد از جمله انجمن سرطان (ACOR) (Meier و همکاران، ۲۰۰۷) که مجموعه ای از ۱۵۹ بخش به هم تنیده است که به اشکال خاصی مربوط به سرطان است. هدف از این تالار های گفتمان، ارائه پشتیبانی و اطلاعات می باشد و اغلب در پژوهش ها نیز استفاده می شوند. در یک تالار گفتمان با ترافیک بالا ممکن است اطلاعات گسترده ای از بیماران وجود داشته باشد که از سالیان دراز ذخیره شده باشند که متأسفانه دسترسی به این اطلاعات بسیار دشوار است. بسیاری از وب سایت هایی که میزبانی این تالارها را به عهده دارند، تنها دارای قابلیت جستجوی ابتدایی هستند (Chun و همکاران، ۲۰۱۲).

۲-۶. روند های جستجو (Google trends)

یکی از منابعی جالب که در مقالات متعدد به آن اشاره شده رونمایی جستجو ها از طریق موتور جستجو گوگل است (<http://www.google.com>). این ابزار، اطلاعاتی در رابطه با تعداد دفعات جستجوی کلید واژه های مختلف می دهد که می تواند با مقایسه با کلید واژه های دیگر مقدار شیوع، زمان شیوع، تعدد شیوع و موارد مختلف دیگر را بدست آورد. به تازگی، گوگل TM با گرفتن کلمات کلیدی پرس و جو به شناسایی جستجو های خاص شامل عبارات ILI پرداخته است (Ginsberg و همکاران، ۲۰۰۹).

با توجه به مطالعات متعدد (Velasquez و همکاران، ۲۰۰۸)، نشان داده شد جستجوهای رایج در وب را می توان به سه گروه، اطلاعاتی، ناوبری و تبدیلی گروه بندی کرد.

- جستجوهای اطلاعاتی: این نوع جستجو به دنبال اطلاعات کلی در یک موضوع گسترده است که به طور معمول یک صفحه وب شامل همه اطلاعات مورد نیاز نیست. کاربران با جستجوهای اطلاعاتی معمولاً سعی می کنند اطلاعاتی از صفحات وب مختلف بدست بیاورند.

- جستجوی ناوبری: که در آن کاربر به دنبال وب سایت هایی با یک اطلاعات واحد است که کاربر در نظر گرفته است. این نوع از کاربران انتظار دارند نتیجه جستجو به سرعت به صفحه ای که اومی خواست، برسد. کاربر به صفحات دیگر که حاوی اطلاعات مربوط به آن اطلاعات نیست، علاقمند نیست.

- جستجوی تبدیلی (تعاملی): که جستجو برای محیط های تبادل در وب مانند خرید یک محصول، دانلود یک فایل، یا رزرواسیون است (Ting و همکاران، ۲۰۱۳). در چنین مواردی، کاربر از موتور جستجو انتظار ارائه لیست از صفحات خدماتی که او می خواهد دارد.

۷-۲. منابع کاربر محور

این منابع از طریق اطلاعاتی که خود کاربران در مورد سلامت خود یا اطرافیان می دهند، تحلیل ها و پیش بینی هایی را انجام می دهد که اعتبار این پیش بینی ها وابسته به تعداد کاربرانی است که اطلاعات خود را در سایت ثبت کرده اند.

نمونه های این نوع سایت ها که در روند های تجزیه و تحلیل استفاده می شوند مختلف است که در مقالات بررسی شده، به دو نوع اصلی آن اشاره شده است (Kamel Boulos و همکاران، ۲۰۱۰):

۱-۷-۲. چه کسی بیمار است: نمایش قدرت نقشه های تولید شده توسط کاربر

چه کسی بیمار است (<http://whoissick.org>) در سال ۲۰۰۶ به عنوان یک سرویس مکان آگاه برای ردیابی و نظارت بر بیماری، با هدف ارائه اطلاعات بیماران فعلی و مکان آنها راه اندازی شد. البته کیفیت اطلاعات همواره وابسته به دقت جزئیات کاربر گزارش دهنده و تعداد کافی گزارش دهندگان (که در واقع بیمار هستند) است.

۲-۷-۲. *We Feel Fine*: تشخیص متون آموزنده و عاطفی در وب اجتماعی

در سال ۲۰۰۶، سایت <http://www.wefeelfine.org> که اپلت مبتنی بر وب و API (رابط برنامه نویسی) است ایجاد شد. این سایت می تواند برای ارزیابی اثرات نوسان رویدادهای جهان واقعی و عواملی مانند شاخص سهام و یا آب و هوا بر روی احساسات و عواطف مردم استفاده می شود و اطلاعاتی در مورد خلق و خوی عمومی جمعیت در تاریخ و مکان های مختلف در سراسر جهان ارائه دهد که اطلاعاتی در مورد سن، جنس، کشور، دولت، و شهر محل اقامت از کاربر می گیرد.

۵-۲. خلاصه فصل

در حالی که ۱۰ سال پیش وب به عنوان موتور تجارت الکترونیک بوده است، در ۵ سال گذشته با افزایش تعامل کاربران و ایجاد محیط های ایجاد محتوا که افراد بیشتری درگیر در آن هستند، افراد در زندگی آنلاین اجتماعی و افزایش اشتراک دانش قرار گرفته اند (Luque و همکاران، ۲۰۱۱). این امر به ویژه برای افرادی که علاقه مند به پیدا کردن افرادی با شرایط بهداشتی مانند خود هستند و به دنبال اطلاعات بهداشتی هستند مهم است. مصرف کنندگان بهداشت اجتماعی، برای اطلاعات سلامت جستجو می کنند (Kummervold و همکاران، ۲۰۰۸؛ Fox و همکاران، ۲۰۰۶)، و ایجاد محتوا در مورد سلامت خود را در پروفایل کاربر، وبلاگ ها، و یا فیلم ها انجام می دهند (Fox & Jones، ۲۰۰۹). اشتراک گذاری تجربیات و دانش می

تواند فراتر از محتوای وب رایج باشد و شامل داده های سلامت ساخت یافته در سایت هایی مانند PatientsLikeMe (Frost & Massagli، ۲۰۰۸) و 23andMe (2009) باشد.

در فرهنگ سریع امروزه، اینترنت به یک منبع مشترک و عمومی برای اطلاعات پزشکی تبدیل شده است (Andreassen و همکاران، ۲۰۰۷؛ Eysenbach & Kohler، ۲۰۰۳، Baker و همکاران، ۲۰۰۳، Hesse و همکاران ۲۰۰۵). نتایج نظر سنجی از Pew Internet & American Life Project نشان داد که ۸۰٪ از کاربران اینترنت اطلاعات بهداشتی خود را از اینترنت بدست می آورند (۲۰۰۶). پزشکان همچنین از اینترنت برای جستجوی اطلاعات و برقراری ارتباط با بیماران استفاده می کنند (Gaster و همکاران، ۲۰۰۳). به علاوه برای ارائه اطلاعات به بیماران و ارائه دهندگان مراقبت های بهداشتی، اینترنت یک رسانه مهم می باشد (Logan، ۲۰۰۴؛ Tian و همکاران، ۲۰۰۹) و این امر باعث شده است که اینترنت به یک رسانه مهم و حیاتی برای پزشکان، دست اندرکاران بهداشت عمومی، و افراد عادی که به دنبال اطلاعات بهداشتی هستند تبدیل شده است. اطلاعات در مورد بیماری ها و شیوع بیماری ها نه تنها از طریق اطلاعیه ها آنلاین سازمان های دولتی بلکه از طریق کانال های غیر رسمی، اعم از گزارش های مطبوعاتی، وبلاگ ها، چت روم ها و جستجوهای وب منتشر می شود. در مجموع، این منابع باعث ارائه دیدی در بهداشت عمومی می شود که اساساً متفاوت از گزارش های بیماری است که توسط زیرساخت های رایج بهداشت عمومی ارائه می شود. در طول ۱۵ سال گذشته، تکنولوژی اینترنت به ابزاری جهت نظارت بر بهداشت عمومی تبدیل شده است. این سیستم با استفاده از اطلاعات غیر رسمی الکترونیکی، باعث کاهش زمان تشخیص شیوع بیماری شده است که باعث پیشگیری دولت ها از شیوع بیماری، و تسهیل پاسخ دهی بهداشت عمومی به شیوع بیماری و بیماری های نوظهور است (Brownstein و همکاران، ۲۰۰۹). از آنجا که منابع مبتنی بر وب اغلب حاوی اطلاعاتی است که از طریق کانال های ارتباطی رایج دولتی بدست نیامده است و مبتنی بر منابع نظارتی روزانه است، در نتیجه برای سازمان های بهداشت عمومی در هشدار شیوع عمومی مفید هستند.

در این فصل به منابع مختلف اطلاعات سلامت در وب پرداخته شده و انواع مختلف آن و اهمیت آنها بیان شد.

فصل سوم - استخراج اطلاعات سلامت از وب

۳-۱. مقدمه

در این فصل به بررسی چگونگی اتصال داده های پراکنده مرتبط با سلامت از جوامع مختلف وب و ارائه دانش یکپارچه اطلاعات سلامت، به طور خاص، داده های رسانه های اجتماعی مبتنی بر جوامع بیماران، سایت های مراقبت درمانی با محتوای تخصصی پرداخته می شود.

در حال حاضر، هیچ مانعی و استاندارد برای انتشار اطلاعات و اعتبار سنجی اطلاعات در WWW وجود ندارد (Craan و Oleske، ۲۰۰۲). اطلاعات بهداشتی موجود در وب هنوز فاقد اعتبار سنجی مناسب هستند. بنابراین سادگی پیدا کردن اطلاعات از وب، هیچ تضمینی در مورد صحت و اعتبار آنها را فراهم نمی کند. با این حال، اطلاعات بهداشتی صحیح از اهمیت زیادی برخوردار است زیرا دانش و پیش بینی غیر قابل اعتماد می تواند بالقوه در سلامت و حتی در زندگی خطرناک باشد (Ku و همکاران، ۲۰۰۸).

برای استخراج اطلاعات سلامت از محیط وب از وب کاوی که مبتنی بر متن کاوی است، استفاده می کنند. متن کاوی فرآیند استخراج اطلاعات با کیفیت بالا از متون زبان طبیعی است که به طور عمده از داده های بدون ساختار یا نیمه ساختار یافته استفاده می کند (Antonio do Prado و Ferneda، ۲۰۰۸). با رشد سریع استفاده از اینترنت، اطلاعات متنی اینترنت با توجه به پیشرفت در تبدیل اسناد کاغذی به قالب الکترونیکی رو به افزایش است. علاوه بر این، در دنیای اینترنت، متن به شایع ترین روش برای تبادل اطلاعات رسمی تبدیل شده است (Berry، ۲۰۰۴). بنابراین، وب کاوی، استخراج الگوهای معنی دار از حجم عظیمی از اطلاعات را بر روی اینترنت، از طریق تکنیک های متن کاوی می گویند (Velasquez و همکاران، ۲۰۰۸). وب کاوی در حال حاضر به دلیل گسترش سریع WWW که اینترنت را تبدیل به یک منبع عظیمی از دانش و اطلاعات کرده است، به حوزه های مختلفی توجه می کند. علاوه بر این، وب کاوی

به عنوان مبنای اصلی ای بر مبنای ابرداده معنایی برای مشخص کردن اسناد و ایجاد دید کلی از محتوای اسناد وب است (Wu و Song، ۲۰۰۹).

محتواکاوی وب (Web Content Mining)، یکی از سه شاخه وبکاوی است که در واقع، کشف اطلاعات مفید از مستندات و داده های ساختیافته و نیمه ساختیافته و غیر ساختیافته وب می باشد. یک شاخه دیگر این مقوله، ساختارکاوی (Web Structure Mining) وب است که به کشف مدل پشت زمینه حاکم بر ساختار فرا پیوند های وب می پردازد و هدف آن، ایجاد اطلاعاتی همچون تشابه یا ارتباط بین سایت های مختلف وب است. شاخه دیگر آن کاربرد کاوی وب می باشد که سعی می کند از تعاملات کاربر با وب، اطلاعاتی کسب کند و از آن ها بصورت سابقه ای در مراجعات بعدی کاربر سود ببرد. در زمینه محتواکاوی وب نرم افزارهای خزنده (Crawler)، به گشت و گذار در اقیانوس وب پرداخته، اقدام به نمایه سازی واژگان در پایگاه داده خود می نمایند که مورد استفاده موتورهای کاوش، در زمان جستجوهای کاربر قرار می گیرد. نمونه بارز این روش، موتور کاوشگر Google است.

در همین راستا ابزارهایی همچون FASTUS: Finite-State Automaton Text Understanding System، در خلال این مأموریت به تجزیه و تحلیل متون، با هدف کشف گروه های مختلف واژگان مانند اسامی، افعال، ترکیبات وصفی و اضاف و غیره می پردازند که به کشف دانش از محتویات وب کمک می کند. این روش هم اکنون برای زبان های انگلیسی و ژاپنی پیاده سازی شده است و بصورت بالقوه برای دیگر زبان ها قابل استفاده است (شهیدی، ۱۳۸۴).

از طرف دیگر استفاده از آنتولوژی (Ontology) در وب در بهینه سازی کاوش در وب پیشنهاد می گردد. آنتولوژی، یک فرهنگ واژگان مشترک بر اساس موضوع سایت برای استاندارد سازی ارائه مفاهیم آن جهت قابل تفسیر شدن توسط ماشین، تعریف می کند. آنتولوژی، یک جزء کلیدی وب مفهومی (Semantic Web) است.

هدف داده کاوی کشف انگاره های معتبر، جدید و قابل ردیابی در حجم عظیمی از داده ها با استفاده از ابزارهای آماری و هوش مصنوعی است. تاریخچه کشف دانش در پایگاه های اطلاعاتی، که امروزه به داده کاوی مشهور است، قدمت چندانی ندارد. در اوایل دهه ۱۹۹۰، هنگامی که اصطلاح کشف دانش در پایگاه های اطلاعاتی برای اولین بار مطرح شد، هجده همگانی به سمت طراحی الگوریتم های داده کاوی صورت پذیرفت و این زمانی بود که شرکت ها اقدام به ذخیره سازی مقادیر عظیم داده ها کرده بودند و به دنبال روش هایی برای بهره وری از این انبار داده ها بودند. در این موعده بود که مفهوم داده کاوی و ابزارهای آن مورد توجه قرار گرفت.

رسالت اصلی داده کاوی در دو طبقه کلی جای می گیرد: توصیف و پیش بینی. در سطح توصیف، هدف فهمیدن داده های زمان گذشته و حال است. از الگوهای توصیف برای جستجوی گروهی از متغیرهای مشابه

در افراد یا دسته‌هایی از گروه‌های جمعیت شناختی مشترک (که ویژگی‌های خاصی از خود نشان می‌دهند) استفاده می‌شود. از پیش‌بینی نیز برای اظهار نظر درباره امور ناشناخته بر اساس امور شناخته شده استفاده می‌شود. از این ویژگی می‌توان برای پیش‌بینی آینده و یا اظهار نظر درباره حال استفاده کرد. در پیش‌بینی دو نوع کارکرد وجود دارد: رده‌بندی (که هدف از آن قرار دادن یک آیتم در یک طبقه است)، و تخمین (که هدف از آن تولید مقادیر عددی برای یک متغیر ناشناخته است).

رایانه‌ها به هنگام ظهور این وعده را دادند که به عنوان یک مخزن دانش و خرد باشند، اما در عوض حجم عظیمی از داده‌ها را به سوی ما روانه ساختند. از طرف دیگر، همگانی بودن اینترنت باعث افزایش تصاعدی مدارک متنی (صفحات وبی، اخبار، مقالات علمی، و ...) شده است. این نوع انفجار اطلاعات چالشی برای نظام‌های بازیابی اطلاعات از نظر مدیریت و بازیابی کارا و اثربخش اطلاعات ایجاد کرده است. فرمول بندی استاندارد یک نیاز اطلاعاتی مستلزم بیان آن در قالب عبارت پرس و جو است؛ به عبارت دیگر، پرس و جو عبارت است از بیان نیاز اطلاعاتی کاربر. با این وجود، بسیاری اوقات فرمول بندی چنین پرس و جویی دشوار است. به عنوان مثال فرض کنید یک کاربر می‌خواهد بداند که در یک ماه معین (و یا در یک محل معین) چه اتفاقاتی رخ داده است. در این مورد، توصیف نیاز اطلاعاتی در قالب یک موضوع منفرد بسیار مبهم خواهد بود زیرا کاربر نمی‌داند که موضوع مورد نظرش چیست. در واقع، کاربر ممکن است تمایل داشته باشد که مجموعه‌ای از موضوعات کشف شده را از مجموعه مدارک موجود استخراج نماید.

وب‌کاوی به استفاده از فنون داده‌کاوی در راستای بازیابی، استخراج، و ارزیابی خودکار اطلاعات به منظور کشف دانش از مدارک و خدمات وبی اشاره دارد. داده‌های وبی غالباً توزیعی، ناهمگن، و بدون برچسب گذاری هستند. بنابراین، هر رابط انسانی نیاز دارد که سوالات مبهم را بررسی کند. تقریباً نود درصد از داده‌ها بی‌استفاده بوده و غالباً هیچ نوع اطلاعات مربوط و مورد نیاز کاربر را ارائه نمی‌دهند. با در نظر گرفتن حجم زیاد ذخیره‌سازی و آمایش داده‌های مورد نیاز برای یک پرس و جوی ساده، پردازش ضرورتاً مستلزم ابزارهای مناسبی است که صرفاً دانش مربوط را به عنوان نتیجه نهایی یافته‌ها ارائه نماید.

جنبه‌هایی نظیر دسته‌بندی، استخراج قوانین وابستگی، ناوبری، شخصی‌سازی، وب‌معنایی، بازیابی اطلاعات، متن‌کاوی و شکل‌کاوی همگی زیرمجموعه وب‌کاوی به شمار می‌روند از این رو، فناوری‌های وب‌کاوی تا حدودی همان فناوری‌های داده‌کاوی می‌باشند، لکن از آنجا که تفاوت‌های ظریفی بین داده‌کاوی و وب‌کاوی وجود دارد استفاده از ابزارها و الگوریتم‌های جدید و اصلاح شده برای مدیریت مناسب اینترنت ضروری به نظر می‌رسد. وب‌کاوی نیز عموماً به داده‌های ساختارنیافته یا شبه-ساختارنیافته وبی می‌پردازد.

وب‌کاوی را می‌توان به طور کلی به سه طبقه تقسیم نمود (خاصه، ۱۳۸۸):

- وب کاوی محتوای مدارک چندرسانه ای، شامل اطلاعات متنی، فرامتن، تصاویر، و دیداری - شنیداری. این طبقه به استخراج سلسله مراتب های مفهومی از وب و دسته بندی خودکار آنها می پردازد.
- وب کاوی ساختار پیوندهای درون مدارک، که نموداری از پیوندهای موجود در یک سایت یا در بین سایت ها تهیه می کند. به عنوان مثال در گوگل، یک صفحه وبی در صورتی مهم تلقی خواهد شد که صفحات مهم دیگر به آن پیوند داده باشند.
- وب کاوی استفاده از داده های تولید شده توسط تعاملات کاربران با وب، که عموماً در قالب گزارش های دسترسی وب سرور، مشخصات کاربران، پرسش های کاربران، و اطلاعات مربوط به کلیک های آنها نشان داده می شود. این طبقه تحلیل گرایش و تحلیل انگاره دسترسی به وب را فراهم می کند.
- همانطور که اشاره شد با استفاده از وب کاوی اطلاعات مفیدی از حجم عظیم اطلاعات وب استخراج می شود. در ادامه به بررسی مدل های مختلف ارائه شده در مقالات در حوزه سلامت در وب، پرداخته می شود.

۲-۳. سیستم یکپارچه زبانی (UMLS)

رویکرد این مدل بر اساس استخراج مفهوم پزشکی با استفاده از سیستم یکپارچه زبان پزشکی (UMLS)، چارچوب شرح منابع (RDF) و مدل سازی معنایی برای نشان دادن سلامت اجتماعی و تجارب پزشکی، و خلاصه سازی اطلاعات یکپارچه در سلامت است (Chun و همکاران، ۲۰۱۲).

این سیستم در حال حاضر، سیستم اطلاعات از PubMed، WebMD، PatientsLikeMe و PubMed را یکپارچه می کند و می تواند برای کسب اطلاعات از طیف گسترده ای از وبلاگ ها مبتنی بر متن مورد استفاده قرار گیرد. این سیستم با استفاده از استخراج اطلاعات مبتنی بر هستی شناسی (ontology) و مدل سازی معنایی از داده های سلامت اجتماعی، اطلاعات غیررسمی مشخصی را یکپارچه می کند که به نوعی از محتوای نوشته شده توسط بیماران است.

با این حال، تنوع محتوا در اینترنت است که اغلب گیج کننده است. اطلاعات در میان بسیاری از سایت ها و در فرمت های مختلف پراکنده شده است. به طور خاص، پیدا کردن اطلاعاتی که در سایت های شبکه های اجتماعی و جوامع بحث آنلاین، که در آن تجارب بیمار به اشتراک گذاشته شده است، بسیار دشوار است. این تجربیات می تواند منابع ارزشمند دانش برای بیماران دیگر، سازمان های دولتی و پزشکان باشد.

هدف در این مدل پیوند داده های پراکنده مرتبط با سلامت از جوامع مختلف وب، و ارائه یک دید یکپارچه از اطلاعات در مورد سلامتی در حالات خاص، علائم، درمان، و تاثیراتی جانبی، برای حمایت از تصمیم گیری های بهداشتی است. همانطور که در مقاله اشاره شده این کار بخشی از پروژه بزرگتری است که قصد

دارد به توسعه کاوش بهینه شده و قابلیت یکپارچه سازی برای سایت های مبتنی بر بحث سلامت و یا سایت های مرتبط با سلامت و همکاری اجتماعی است پردازد، که موارد زیر را فراهم می کند:

- جستجو بر اساس مشخصات که در آن ویژگی های پزشکی کاربر برای پیدا کردن اطلاعات تولید شده توسط بیماران مشابه مورد استفاده قرار می گیرد.

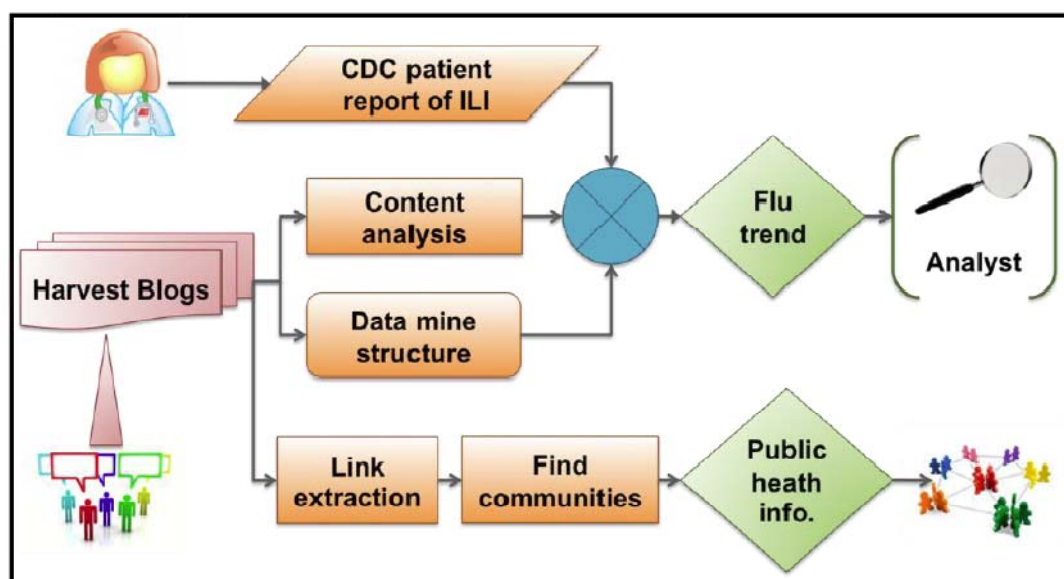
- اطلاعات ارائه شده به کاربر به صورت یکپارچه و قالب درخواستی که بر اساس ویژگی های پزشکی کاربر، فیلتر شده است.

برای جلوگیری از مسائل حفظ حریم خصوصی، این مدل بر اطلاعات موجود بر روی وب باز متمرکز شده است، به عنوان مثال، سایت های که توسط کلمه عبور محافظت نمی شوند. کاربران به عنوان بیمار، مراقب یا افراد دیگر بدون زمینه پزشکی که ممکن است در جستجوی اطلاعات مربوط به وضعیت سلامت خود باشند، فرض می شوند.

۳-۳. داده کاوی رسانه های اجتماعی (WSM)

یکی از استراتژی های پاسخ مناسب به شیوع استفاده از WSM برای کاهش جمعیت و تاثیر بیماری بر انسان می باشد (Corley و همکاران، ۲۰۱۰).

چارچوب پیشنهادی این مدل، در شکل ۳-۱، چارچوب این مدل توضیح داده شده که برای شناسایی و پاسخ به شیوع آنفلونزا و بیماری های مشابه است. همانطور که نشان داده شده، در این مدل اطلاعات از دو منبع یعنی اطلاعات مبتنی بر وب مانند وبلاگ ها و منابع دولتی (CDC) دریافت شده و با مقایسه آنها با هم به پیش بینی روند بیماری پرداخته می شود.



شکل ۳-۱. روش نظارت بر بیماری های شبیه آنفلونزا از طریق رسانه های اجتماعی (Corley و همکاران، ۲۰۱۰)

۳-۴. استخراج کننده اطلاعات (IE)

موضوع اصلی در مدل IE، کیفیت نتایج است (Etter و همکاران، ۲۰۱۰). دو نکته اصلی در این مورد وجود دارد. اول، تشخیص معیارهای کیفیت: معیارهای عینی که صحت اندازه گیری و تجزیه و تحلیل سیستم در شرایط عادی را اندازه می گیرد (F-اندازه گیری، جامعیت و مانعیت)، و از سوی دیگر، معیارهای موضوعی است که اندازه گیری ارزش نتایج توسط کاربر نهایی است.

در این مدل برای مثال MedISys مقالات خبری را از هزاران منبع بر خط در سراسر جهان می یابد که شناسایی مقالات بالقوه به نظارت اپیدمی مرتبط است و با استفاده از یک کلمه کلیدی مبتنی بر جستجوی وب اطلاعات را از طریق خوراک RSS به PULS به طور مستمر می فرستد. سپس PULS یک حقیقت یاب به کار برده است که بر اساس استخراج اطلاعات (IE)، دقیقاً آن چیزی که در مقاله رخ داده است را شناسایی می کند. مثلاً چه کسی بیمار است و در چه شرایط و مکانی قرار دارد. سپس اطلاعات در پایگاه داده ذخیره می شود. مطالبی که باعث ایجاد یک رکورد در پایگاه داده نشوند دور انداخته می شوند. و در مرحله آخر به تعیین ارتباط مطالب انتخاب شده با موارد توصیف شده در حوزه بهداشت عمومی یا به طور خاص برای نظارت اپیدمی پرداخته می شود. همانطور که اشاره شده مهم ترین نکته در این مدل کیفیت است.

۳-۴-۱. معیارهای کیفیت

در این بخش به یک دیدگاه دقیق در دید رایج از کیفیت در تجزیه و تحلیل متن به طور کلی، و به طور خاص IE پرداخته می شود و به سوالات زیر پاسخ داده می شود:

- چه تعریف کیفیت مناسبی برای استفاده در مدل وجود دارد؟

- چگونه باید کیفیت را اندازه گیری کرد؟

طبقه بندی زیر برای کیفیت پیشنهاد می شود:

• کیفیت بر مبنای هدف: چشم انداز سیستم

که بر مبنای ۲ عامل صحت (Accuracy) و ارتباط درونی (Inconsistency) است.

• کیفیت بر مبنای موضوعی: چشم انداز کاربر

که بر مبنای دو عامل سودمندی (Utility) یا ارتباط با نیاز و قابلیت اطمینان (Reliability) است.

در سطح بالا، بین اهداف در مقابل موضوع باید تمایز قائل شد. بسیاری از تحقیقات در IE در دو دهه گذشته بر روی صحت است (Hirschman، ۱۹۹۸؛ ACE، ۲۰۰۴). صحت اندازه گیری معیاری است که نشان دهنده دقت استخراج معانی از یک مقاله یا متن است.

ارتباط درونی نیز برای برآورد احتمال صحت پاسخ سیستم مورد مطالعه قرار گرفته است (Culotta و McCallum، ۲۰۰۴).

در سمت موضوعی، کارآمدی و مرتبط بودن نشان دهنده مفید بودن برای کاربر است. در اینجا چندین نکته وجود دارد. اول، آن است که این اندازه گیری یک اندازه گیری کاملاً موضوعی است، و به آسانی در یک حالت خاص قابل محاسبه نیست. دوم دیدگاه کاربر است که چقدر اندازه گیری را صحیح بداند. برای مثال، یک مورد استخراج شده می تواند ۱۰۰ درصد درست و در عین حال بسیار کم ارزش برای کاربر باشد (در نظارت همگانی) مثلاً یک رویدادی واقعاً استخراج شده است که در گذشته دور اتفاق افتاده که در حال حاضر مهم نیست. در مقابل، اطلاعاتی با دقت پایین استخراج شده ممکن است اهمیت و ارزش فراوانی برای کاربر داشته باشد.

در نهایت، معیار مطمئنی که نشان می دهد که آیا این رویداد یک گزارش "واقعی" است مثلاً ارزش اطلاعات استخراج شده ممکن است بالا باشد، اما آیا درست است؟ آیل اطلاعات قابل اعتماد است؟ از آنجا که هدف نهایی از هر گونه فرایند نظارت، کیفیت است، لیست کاملی از معیارهای کیفیت بیان شد. با این حال، پاسخ به این نیاز به مقدار زیادی از دانش خارجی به سیستم، که تنها می تواند توسط کاربر انسانی از طریق یک فرآیند راستی آزمایی تست شود به دست می آید. البته اندازه گیری کیفیت فراتر از محدوده این مقاله است.

۳-۵. سیستم بازیابی اطلاعات وب (WebIRS)

سیستم بازیابی اطلاعات وب (WebIRS)، برای کمک به متخصصان مراقبت های بهداشتی برای به دست آوردن اطلاعات به روز دانش پزشکی و اطلاعات از طریق شبکه جهانی وب (WWW) طراحی شده است (Ting و همکاران، ۲۰۱۳).

دانش سنتز و سازماندهی اطلاعاتی است که در توضیح وقایع و شرایط کمک کننده است. دانش همچنین کمک می کند تا به ارزیابی و ایجاد ارتباط بین اتفاقات حال حاضر و آینده رسید، بنابراین به عنوان تسهیل کننده تصمیم گیری است (Housel و Bell، ۲۰۰۱). به طور خاص در زمینه های پزشکی، دانش برای قضاوت و تصمیم گیری بالینی به متخصصان پزشکی مهم است. بنابراین به عنوان کمکی به بهبود کیفیت مراقبت ها و خدمات ارائه شده به بیماران است.

بازیابی اطلاعات به معنای پیدا کردن مواردی است که عمدتاً اسناد بدون ساختار (معمولاً متن) هستند که نیاز اطلاعاتی را از مجموعه بزرگ اطلاعات برآورده می کنند (Manning و همکاران، ۲۰۰۸) و برای جستجوی اسناد، اطلاعات، ابرداده ها در WWW است (Vakali و Pallis، ۲۰۰۷). با توجه به نفوذ روز افزون فن آوری و WWW، میلیون ها نفر از مردم با بازیابی اطلاعات به صورت روزانه و زمانی که آنها با استفاده از یک موتور جستجو می کنند و یا حتی در ایمیل خود جستجو می کنند، در تعامل هستند. بازیابی اطلاعات امروزه به شکل غالب دسترسی به اطلاعات تبدیل شده است. فرایند اصلی که در این مدل استفاده شده است در شکل ۲-۳ نشان داده شده که به شرح زیر است:

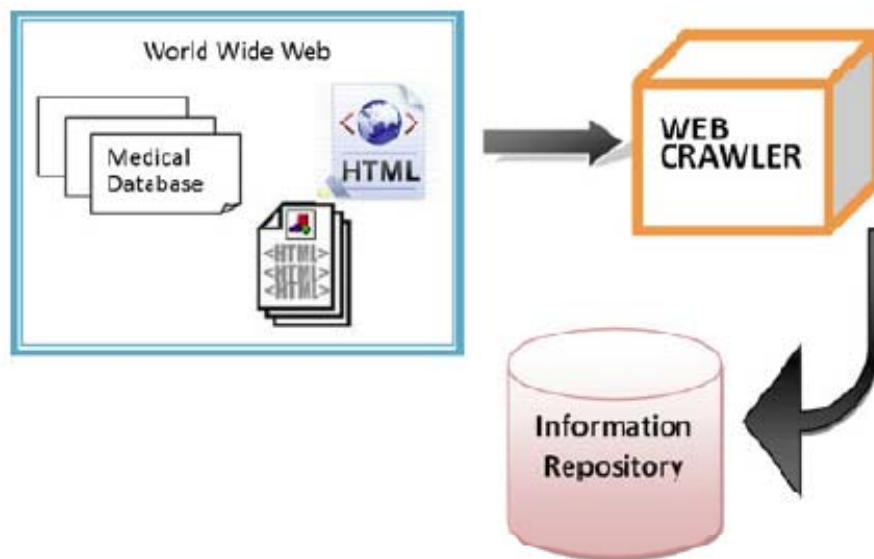


Fig. 2 Web crawling process

شکل ۲-۳. فرایند خزش در وب (Ting و همکاران، ۲۰۱۳)

۱. ماژول آماده سازی و بازیابی

(۱) خزنده وب

(۲) پیش پردازش

i. حذف برچسب های HTML

ii. حذف واژه های توقف (Stop Key)

iii. ریشه یابی

(۳) انتخاب کلید واژه و تعیین وزن

۲. ماژول طبقه بندی سند

(۱) الگوریتم های طبقه بندی کننده بیز

۳. ماژول خلاصه متن

(۱) استخراج جملات کلیدی

(۲) تنظیم جملات اصلی

(۳) ایجاد خلاصه

۳-۶. نمونه گیری

سکوهایی مانند توییتر برای محققان فرصت کافی برای تحلیل پدیده های اجتماعی فراهم آورده است. با این حال چالش قابل توجه در این زمینه محاسبات زیاد با توجه به میزان زیاد تولید اطلاعات جدید است (Choudhury و همکاران، ۲۰۱۰). بنابراین محققان اغلب مجبور به تجزیه و تحلیل یک "نمونه (sample)" که انتخاب عادلانه از داده ها هستند، هستند. مانند دیگر پدیده های رسانه های اجتماعی، انتشار اطلاعات یک فرایند اجتماعی است که تحت تاثیر کاربر، موضوع، و توپولوژی گراف توزیع است. در طول چهل سال گذشته، روش های سنتی مطالعه فرایندهای اجتماعی مانند نفوذ اطلاعات، شناسایی تخصصی و تشخیص جامعه به صورت محلی متمرکز شده است و مطالعات بر گروه های نسبتاً کوچک انجام شده است. با این حال، گسترش گسترده چندین وب سایت اجتماعی مانند فیس بوک، توییتر، آواتار، فلیکر و یوتیوب راههای فراوانی به محققان برای مطالعه چنین فرآیندهایی در مقیاس بسیار بزرگ ارائه داده است. این امر به دلیل این است که داده های الکترونیکی اجتماعی را می توان در طول فواصل زمان طولانی، و برای جمعیت های بسیار بزرگ به دست آورد و ذخیره کرد. نتیجه این است که مطالعه در فرایندهای اجتماعی در مقیاس میلیون گره، که یک دهه پیش غیر قابل تصور بوده است، در حال تبدیل شدن به روال عادی است.

در نتیجه محققان، امروزه قادر به مطالعات تجربی هدایت شده در انتشار اطلاعات هستند. از جمله مطالبی که در وبلاگ (Gruhl و همکاران، ۲۰۰۴)، داده های زنجیره ای در اینترنت (Liben-Nowell و Kleiberg، ۲۰۰۸)، برچسب گذاری اجتماعی (Anagnostopoulos و همکاران، ۲۰۰۸)، فیس بوک خبری (Sun و همکاران، ۲۰۰۹)، بازی های آنلاین (Bakshy و همکاران، ۲۰۰۹) و موارد دیگر، منتشر می شود.

با این حال توجه به حجم داده ها، کمتر مورد توجه است اما دو چالش به همان اندازه مهم است:

۱. چالش اکتساب داده ها و پیچیدگی اطلاعات تجزیه و تحلیل شده. به عنوان مثال، در شبکه اجتماعی فیس بوک در حال حاضر تعداد کاربران بیش از ۳۵۰ میلیون کاربر است و رسانه اجتماعی توییتر نرخ حدود ۱۷،۰۰۰ پست (توییت) در دقیقه را دارد. در چنین شرایطی، ابتدا، ابزارهای کسب

داده ها که به طور معمول رابط های برنامه کاربردی است (API) و در دسترس عموم است اغلب برای پیگیری تمام داده هایی که تولید می شوند کافی نیست. توجه داشته باشید، حد نرخ تماس API برای تویتر فقط ۲۰,۰۰۰ توییت در ساعت است. از این رو ایجاد یک گلوگاه کسب اطلاعات بسیار مهم است.

۲. هزینه منابع گسترده ای که در ذخیره سازی داده ها در این مقیاس استفاده می شود بسیار بالاست و نیز پیچیدگی بطور قابل توجهی بالا در تجزیه و تحلیل داده ها وجود دارد.

رویکرد نمونه گیری شامل دو مرحله است. اول، استفاده از چندین تکنیک نمونه گیری رایج از قبیل نمونه گیری تصادفی، نمونه گیری مبتنی بر درجه فعالیت کاربر، آتش در جنگل (forest-fire) و نمونه برداری بر اساس مکان که برای استخراج زیر گراف هایی از یک گراف اجتماعی از کاربران درگیر در فعالیت های اجتماعی است. دوم، این زیر گراف ها در مطالعه ویژگی های انتشار با توجه به خواص کاربران (به عنوان مثال مشارکت)، ساختار (به عنوان مثال دسترسی و گسترش) و ویژگی های موقتی (به عنوان مثال نرخ) و همچنین ارتباط با حوادث در جهان خارجی (به عنوان مثال جستجو و روند اخبار) استفاده می شود. برای نمونه گیری همچنین می توان از روش های مبتنی بر فیلتر استفاده کرد. اولین فیلتر، فیلتر کلمات کلیدی با استفاده از یک لیست مربوط به بیماری ها، نشانه ها، و درمان است. (برای مثال لیست wrongdiagnosis.com (Paul و Dredze، ۲۰۱۲). فیلتر بر اساس کلید واژه ها به دلیل اینکه کلمات کلیدی بهداشتی می توانند در بسیاری از زمینه ها استفاده شوند کافی نیست (Culotta، 2010b). در عوض، برای به دست آوردن داده برای طبقه بندی نظارت شده از Mechanical Turk (MTurk) استفاده می شود (Callison-Burch و Dredze، 2010). Turkers توییت ها را چنین برچسب گذاری می کند:

- بیمار: این پیام نشان می دهد که کاربر دارای یک بیماری حاد (به عنوان مثال، سرما، آنفولانزا) است.
- سلامتی: نظر ارسالی به صورت کلی در مورد بهداشت کاربر و یا کس دیگری بود (به عنوان مثال، بیماری های مزمن، شیوه زندگی، رژیم غذایی).

۷-۳. مشکلات زبان فارسی در وب کاوی

زبان فارسی، در تلاقی با جهان الکترونیکی، بخصوص از بعد رسم الخط، دارای مشکلاتی است که کاوش در محتویات آن را دچار کم کیفیتی می نماید (شهیدی، ۱۳۸۴). این مشکلات ریشه در خصوصیات محتوای فارسی منتشر شده در وب است.

۳-۷-۱. خصوصیات وب های فارسی از نظر زبان

عدم وجود یک استاندارد و شناور بودن ویژگیهای رسم الخط و مفاهیم در زبان فارسی، موجب گردیده است تا تقریباً به تعداد صفحات وب فارسی، سبک و سیاق نگارش این زبان بکار رفته باشد. لیکن خصوصیات مشترک اکثر وب های فارسی زبان را می توان چنین ارزیابی نمود:

- الف) نگارش برخی از وب های فارسی، زبان غیررسمی یا محاوره ای است.
 - ب) در وب های فارسی، بخصوص در متون علمی، اغلب واژه های بیگانه، بکرات استفاده می شود که بعضی از آنها بارسم الخط زبان اصلی نوشته می شوند.
 - ج) رسم الخط وب های فارسی، اصولاً غیراستاندارد و متغیر است و اغلب در معرض نوآوری است.
 - ه) نوشته های وب های فارسی، حاوی غلط های تایپی و نگارشی نسبتاً زیادی است، هرچند که اغلب وب های فارسی مهم و پرخواننده، نگارش قابل قبولی دارند.
 - و) رسم الخط وب های فارسی، تابع محدودیت های محیط الکترونیکی و عدم تطبیق آن با الزامات خط فارسی است.
- برای رفع این مشکلات و بهبود استخراج اطلاعات از محتوای وب روش های مختلفی ارائه شده است.

۳-۷-۲. روش هایی برای بهبود کاوش وب های فارسی

- الف) انتخاب مناسب سرعنوان های موضوعی در وب های فارسی
 - ب) استمداد از علم اصطلاح شناسی (Terminology) در نمایه سازی ماشینی: توجه به اصطلاحات و اصطلاح سازی نیز با توجه به ضرورت روزآمد بودن واژگان علمی و تخصصی و لزوم کنترل ورود اصطلاحات بیگانه امری است که ما را ناگزیر به استمداد از علم اصطلاح شناسی و می دارد.
 - ج) تعریف یک استاندارد برای مفاهیم و رسم الخط فارسی در وب
- همانطور که گفته شد، یک تفاوت زبان فارسی با زبان انگلیسی (و زبان های هم ارز)، تنوع املائی یا رسم الخطی کلمات آن است. بعبارت دیگر، در زبان انگلیسی، تنوع در مفهوم کلمات وجود دارد. یعنی برای بعضی مفاهیم، ممکن است کلمات متنوعی استفاده شود. برای مثال کلمات "Hi" و "Hello" دارای مفهوم یکسانی هستند. لیکن در فارسی، علاوه بر وجود کلمات متنوع برای مفاهیم یکسان، مثل "کامپیوتر" و "رایانه"، تنوع در رسم الخط یک کلمه نیز فراوان بچشم می خورد. بعبارت دیگر، در حالی که شما بدنبال صفحات محتوی کلمه "امپراتور" می گردید، کلیه صفحات محتوی کلمه "امپراطور" را از دست می دهید.
- د) استفاده از مفرد و جمع در نمایه سازی
- استفاده از اسامی جنس، نحوه جمع بستن کلمات بصورت باقاعده با بدون قاعده (جمع های مکسر) معضلی است که در نمایه سازی واژگان فارسی معمولاً بسیاری از صفحات وب را شامل نمی شود.

ه) استفاده از یک واسط کاوش فارسی برای رفع چالش های رسم الخط و مفهومی استاندارد سازی رسم الخط فارسی در رایانه ها، ممکن است در ابتدای تولید اولین صفحات وب فارسی بسیار مفیدتر بنظر رسد، لیکن در حال حاضر، با وجود تعداد بسیار زیادی صفحات وب فارسی که در هر حال با عدم نظارت یک استاندارد، تولید شده اند، چندان مثمر ثمر واقع نمی شود. اگر چه ایجاد آن برای تولید صفحات وب فارسی آتی، لازم است. بعبارت دیگر، جهت انجام عملیات وب کاوی در صفحات وب فارسی کنونی، بایستی روشی ابداع کرد تا با توجه به چالش های بحث شده، نتایج مطلوبی از وب کاوی در آنها بدست آید.

۳-۸. خلاصه فصل

با وجود توانایی های بالقوه PML، کشف دانش پزشکی از PML چالش قابل توجهی است. مرحله اول، اطلاعات استاندارد نشده و بدون ساختار است، که از طرف کسانی که آموزش پزشکی ندیده اند منتشر می شود. دوم، صفحات وب همیشه زمان دار نیستند و در نتیجه، تجزیه و تحلیل سری های زمانی را دشوار می کند. در نهایت، PML داده ای علمی نیست اما ممکن است سیگنال های درستی را ارائه کند (MacLean & Seltzer, ۲۰۱۱).

این اطلاعات که به صورت سندهایی در وب وجود دارد، مخلوط جمله هایی است که شامل موضوعات مختلف است و هر موضوع تعریف شده نیز شامل چندین کلمه است. استفاده از روش های مبتنی بر استنتاج و بر اساس موضوعاتی که در آن بیشتر تکرار می شوند باعث ارتباط معنایی قوی در مدل استخراج اطلاعات می شود. البته نتایج در این مدل ها به وضوح نشان بیماری های خاص نیست. برای مثال، بسیاری از موضوعات شامل عبارات در مورد عمل جراحی است، اما معلوم نیست این جراحی ها با بیماری های خاص مانند صدمات جسمی یا سرطان رابطه دارند یا خیر. به علاوه این موضوعات، ما نیاز به مدلی برای شناسایی ساختار ضمنی بیماری داریم (Paul و Dredze, ۲۰۱۲).

در این فصل به بررسی های مدل های مختلف استخراج اطلاعات از وب، معیارهای کیفیت و در نهایت مشکلات زبان فارسی پرداخته شد.

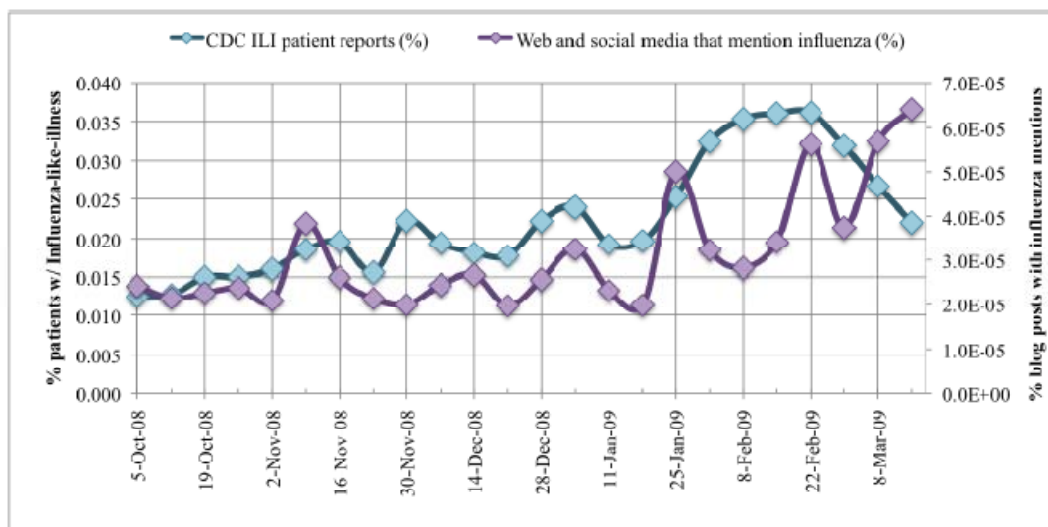
فصل چهارم - پیش بینی سلامت با وب

۴-۱. مقدمه

داده کاوی متن و ساختار وب و رسانه های اجتماعی (WSM) یک منبع نظارتی برای بیماری فراهم می کند و می تواند جوامع آنلاین برای ارتباطات هدفمند بهداشت عمومی (PHC) را شناسایی کند. متن کاوی نشان داد که مطالب ارسالی در مورد آنفلونزا با روند بیماری آنفلونزا در دنیای واقعی یکسان است (Corley و همکاران، ۲۰۱۰). همچنین تحقیقات منتشر شده بر روی نظارت اینترنتی آنفلوانزا که شامل جستجو است (Eysenbach، ۲۰۰۵)، با استفاده از مجموعه ای از search query های یاهو است که شامل کلمات "آنفلوانزا: flu" یا "آنفلوانزا: influenza" (Polgreen و همکاران، ۲۰۰۸) و یا Log های وب سایت های سلامت است (Hulth و همکاران، ۲۰۰۹). منابع اطلاعاتی دیگر، از قبیل خدمات تریاژ تلفنی، می توانند برای تشخیص ILI مفید باشند. یافته ها در Yih و همکاران، (۲۰۰۹) نشان می دهد که خدمات تریاژ تلفنی برای نظارت بر آنفلوانزای با توجه به پوشش خدمات آن قابل اعتماد نیستند، با این حال، ممکن است در شرایط خاصی که در آن سایر اقدامات نظارت ناکافی هستند مفید باشند.

همچنین Greene و همکاران (۲۰۱۰) چگونگی اشتراک گذاری اطلاعات پزشکی را در فیس بوک مطالعه کرده اند، که در آن گروه های خاص اطلاعات بیماری هایی را به اشتراک می گذاشتند و از بیماران حمایت می کردند. Bonander و Karlsen، Fernandez-Luque (۲۰۱۱) برای استخراج اطلاعات از برنامه های کاربردی وب اجتماعی جهت شخصی سازی مراقبت های بهداشتی فرد، رویکرد های مختلفی را بررسی کرده اند.

در شکل ۴-۱ همبستگی روند شیوع بیماری که از طریق ارگان های رسمی اعلام شده است با روند انتشار اطلاعات در محیط های اجتماعی نشان داده شده است که نشان دهنده همبستگی مناسب با هم است.



شکل ۴-۱. ارتباط متون منتشر شده در رسانه های اجتماعی و شیوع بیماری آنفولانزا از ۵ اکتبر تا ۲۱ مارس ۲۰۰۹ (Corley و همکاران، ۲۰۱۰)

به طور خاص در زمینه های پزشکی، دانش برای قضاوت و تصمیم گیری بالینی به متخصصان پزشکی مهم است. بنابراین محیط هایی همچون وب به عنوان کمکی به بهبود کیفیت مراقبت ها و خدمات ارائه شده به بیماران است. (Ting و همکاران، ۲۰۱۳).

متن کاوی و وب کاوی می تواند کاربران را با روش هایی در مقایسه اسناد، رتبه بندی اهمیت و ارتباط آنها و پیدا کردن الگوها و روندها در سراسر اسناد مختلف حمایت کند (Han و Kamber، ۲۰۰۶). متن کاوی و وب کاوی به طور گسترده ای در مسائل مربوط به پشتیبانی از تصمیم گیری در حوزه بهداشت و درمان استفاده شده اند (Ting و همکاران، ۲۰۰۹)، مانند کمک در تدوین اصطلاحنامه پزشکی (Ting و همکاران، ۲۰۰۸)، پیش بینی حمله قلبی (Patil و Kumaraswamy، ۲۰۰۹)، و ارائه اطلاعات قابل اعتماد و جدید به پزشکان. به عنوان مثال، Mostafa و Lam (۲۰۰۰) با استفاده از یادگیری نظارتی به طبقه بندی مدارک پزشکی در حوزه زیست شناسی سلولی پرداخته اند. که از مطالعه آنها مشخص شده است که درجه بالایی از دقت و صحت طبقه بندی با استفاده از روش نظارتی به دست آمده است. Liu و Chu (۲۰۰۷) روش گسترش پرس و جو بر پایه دانش (Knowledge-based) را ارائه کرده اند که از بهره برداری از منبع دانش UMLS به ارائه پاسخ های مربوط به حالات خاصی که در عمل های پزشکی رخ می دهد، می پردازند. Elhadad و همکاران، (۲۰۰۵) یک مدل کاربر واحد برای ایجاد یک خلاصه مناسب از اسناد مربوط به پزشکان با بهره گیری از نظم و قاعده در ساختار متن مقالات پزشکی و محتوا استفاده کرده اند. که از تطبیق کاملی از متن کاوی و وب کاوی در حوزه بهداشت و درمان استفاده کرده است. هدف از این مطالعه استفاده از یادگیری نظارتی برای طبقه بندی اسناد و مدارک پزشکی به دسته های خاص در راستای

تحقق نیازهای کاربر است که به عنوان ارائه کننده اطلاعات پزشکی قابل اعتماد به پزشکان برای دستیابی به اهداف ایشان با استفاده از اطلاعات وب و به منظور ارتقاء کیفیت تصمیم گیری پزشکان، استفاده شود. منابع اطلاعاتی سلامت در وب مزایا و معایبی دارند که در پیش بینی و نظارت بسیار اهمیت دارند. مزایا عبارتند از: کم هزینه بودن و انتقال سریع اطلاعات و معایب عبارتند از: مؤلف کور، عدم ذکر منبع و در نظر گرفتن نظرات به جای حقیقت. مهم ترین کاربرد منابع سلامت در وب همانطور که اشاره شد، نظارت و پیش بینی می باشد که در ادامه به بررسی مفهوم نظارت و روش های پیش بینی پرداخته می شود.

۴-۲. نظارت اپیدمی

نظارت سندرومی، نظارتی است بر سندرمهایی که تاثیر قابل توجهی بر سلامت عمومی، تخصیص منابع، سیاست های بهداشتی پزشکی و آموزش و پرورش می گذارند (Paul و Dredze، ۲۰۱۱). بسیاری از بیماری های شایع با جمع آوری داده ها از مراکز مراقبت های بهداشتی که یک روند شناخته شده است به طور مداوم مانیتور می شوند. منابع، نظارت را محدود می سازد، به خصوص برای بازخورد در زمان واقعی آن. به همین دلیل، وب به یک منبع از نظارت سندرومی در مقیاس وسیع تر و کسری از هزینه تبدیل شده است. Google Flu Trends (Ginsberg و همکاران، ۲۰۰۹) آهنگ نرخ آنفلوانزا را با استفاده از روندیابی پرس و جو های روزانه، تا ۷ تا ۱۰ روز سریع تر از مرکز کنترل بیماری (CDC) نمایش می دهد (Carneiro و Mylonakis، ۲۰۰۹). همبستگی بالا بین آنچه گوگل نمایش می دهد و سایر بیماری ها وجود دارد (Pelat و همکاران، ۲۰۰۹) از جمله در "بیماری لایم (Lyme)" (Seifter و همکاران، ۲۰۱۰). این نتایج تحت قالب infodemiology قرار می گیرند (Eysenbach، ۲۰۰۹). نتایج مشابهی برای تویتر وجود دارد که می تواند منابع مکملی برای روندهای پرس و جو باشد، که البته ممکن است حاوی اطلاعات دقیق تری باشند که مردم آنها را برای خواندن دیگران می نویسند. Lampos و Cristianini (۲۰۱۰) و Culotta (۲۰۱۰) نشان داده اند که توییت ها همبستگی بالایی با داده های تاریخی آنفلوانزا دارند.

به طور مشابه Quincey و Kostkova (۲۰۱۰) توییت ها را در بیماری همه گیر H1N1 برای تجزیه و تحلیل جمع آوری کرده است. Klein و Osborne, Ritterman (۲۰۰۹)، پیش بینی بازارها را با تویتر در پیش بینی H1N1 استفاده کرده اند. به طور کلی، Chew و Eysenbach (۲۰۱۰) تویتر را به عنوان وسیله ای برای نظارت بر درک عمومی از بیماری همه گیر H1N1 ارزیابی کرده اند. Scanfeld و Larson (۲۰۱۰) استفاده نادرست آنتی بیوتیک را از توییت ها نشان داده اند به عنوان مثال استفاده از آنتی

بیوتیک ها برای آنفلوانزا بررسی شده است. جامعه بهداشت عمومی نیز با توجه به چگونگی استفاده از رسانه های اجتماعی در گسترش اطلاعات سلامت، برنامه های کاربردی را توسعه داده است. Vance ، Howe و Dellavalle (۲۰۰۹) تجزیه و تحلیل جوانب مثبت و منفی استفاده از رسانه های اجتماعی برای گسترش اطلاعات بهداشت عمومی در جوانان بررسی کرده اند.

نظارت اپیدمی بر دو نوع شاخص محور و اتفاق محور است (Hartley و همکاران، ۲۰۱۰). در حالی که اطلاعات نظارت سابقا بر اساس داده های کمی ساخت یافته مانند اطلاعاتی از آزمایشگاههای بالینی ملی یا بین المللی و یا پایگاه داده، که دارای کیفیت قابل اعتماد است جمع آوری می شد، در روش های جدید که خطای بیشتری دارد متکی بر هشدار و پایش وضعیت است. این اطلاعات به ویژه از رسانه های منبع باز، مانند سایت های خبری آنلاین دریافت می گردد. در حالی که این منابع اطلاعاتی به طور کلی کمتر قابل اعتماد هستند، با این وجود آنها تشکیل یک کانال حیاتی از اطلاعات در سلامت عمومی داده اند. دلیل آن این است که این رسانه ها شامل موارد بسیار نادر و ضعیف است که ممکن است نشان دهنده ظهور حوادث مهم مانند یک بیماری همه گیر بدوی و یا تغییرات مهم در وضعیت سلامت عمومی باشند و در بسیاری از موارد آنها می توانند بسیار سریع تر از کانال های رسمی عمل کنند. برای مقامات بهداشتی ملی و فرا ملی (به عنوان مثال، در سطح اروپا و آسیا) نیاز به اطلاعات به موقع در مورد تهدید بیماری ها و بیماری های همه گیر عفونی نوظهور برای عموم مردم مطرح است. بنابراین، این سازمان ها در نظارت بر این رسانه ها تکیه می کنند (Etter و همکاران، ۲۰۱۰).

در این نوع سیستم ها، سیستم تصمیم گیری نمی کند، اما از طریق فیلتر کردن حجم عظیمی از اطلاعات و تلاش برای شناسایی موارد خاص، در تصمیم گیری پشتیبانی می کند.

استفاده کنندگان از چنین سیستم هایی متخصصان پزشکی (EIOS) هستند، که به طور کلی در روش های محاسباتی آموزش ندیده اند و بنابراین ابزاری که از آن استفاده میکنند باید محسوس باشد و نباید کاربر را با حجم و پیچیدگی درگیر نماید. مثلا این سیستم ها باید نشان دهند چه کسی بیمار است و در چه شرایط و مکانی قرار دارد. سپس اطلاعات در پایگاه داده ذخیره می شود. مطالبی که باعث ایجاد یک رکورد در پایگاه داده نشوند دور انداخته می شوند. در مرحله آخر به تعیین ارتباط مطالب انتخاب شده با موارد توصیف شده در حوزه بهداشت عمومی یا به طور خاص برای نظارت اپیدمی پرداخته می شود.

۳-۴. تجزیه و تحلیل نفوذ اجتماعی

تجزیه و تحلیل انتشار اطلاعات اجتماعی مورد علاقه برای پژوهشگران حوزه های مختلف اعم از علوم اجتماعی، اپیدمیولوژی، انتشار بیماری، فیزیک و اقتصاد (Zachary ، ۱۹۷۷)، (Newman ، ۲۰۰۲)،

(Watts و Dodds ، ۲۰۰۷) بوده است. قبلا کارهایی برای مدل سازی و پیش بینی مسیر انتشار اطلاعات در شبکه های اجتماعی مفید برای برنامه های مختلفی اعم از سیستم های توصیه، تبلیغات آنلاین، پیش بینی رفتار کاربر و مهار بیماری (Kempe و همکاران، ۲۰۰۳)، (Gruhl و همکاران، ۲۰۰۴)، (Song و همکاران، ۲۰۰۶)، (Anagnostopoulos و همکاران، ۲۰۰۸)، (Kossinets و همکاران، ۲۰۰۸)، (Liben-Nowell و Kleiberg، ۲۰۰۸) ایجاد شده است. در تحقیقات گذشته (Kempe و همکاران، ۲۰۰۳)، نویسنده پیشنهاد راه حلی برای مشکل بهینه سازی انتخاب گره با نفوذ در شبکه های اجتماعی که می تواند به صورت آبشاری فرزندان بیشتری انتخاب شوند داده است. در کارهای اخیر، Bakshy و همکاران (۲۰۰۹) مطالعه چگونگی "حرکات (gestures)" در دنیای آنلاین زندگی یعنی زندگی دوم را بررسی کرده اند. در تحقیق دیگری، Sun و همکاران (۲۰۰۹) الگوهای انتشار در فیس بوک از طریق News Feed را بررسی کرده اند و نتیجه گرفته اند که در رسانه های اجتماعی آنلاین، پویایی از طریق انتشار اطلاعات و اغلب با برخورد زنجیره های کوتاه اطلاعات ایجاد شده است.

اگر چه کارهای قبل بینش مفیدی در ویژگی های انتشار در رسانه های اجتماعی می دهد اما محدودیت هایی در این زمینه وجود دارد (Choudhury و همکاران، ۲۰۱۰):

- محتوای اطلاعات: رسانه های اجتماعی آنلاین امروزه وابسته به محتوای اطلاعات منتشر شده است؛ به عنوان مثال شواهد تاریخی که در ارتباط با حوادث خارجی منتشر می شود (Gruhl و همکاران، ۲۰۰۵). از این رو، نمونه برداری خالص مبتنی بر توپولوژی ممکن است مناسب برای مطالعه فرایندهای اجتماعی که در رابطه بین محتوای به اشتراک گذاشته شده و اقدامات کاربر و رویدادهای خارجی است، نباشد.
- زمینه های اجتماعی: بسیاری از تحقیقات پیشین اطلاعات زمینه ای از کاربران در گراف اجتماعی، مانند موقعیت جغرافیایی، و یا چگونگی تغییر سریع وضعیت کاربر (به عنوان مثال نرخ فعالیت های اجتماعی) را در نظر نمی گیرند. چنین اطلاعاتی زمینه ای در مطالعه پدیده های اجتماعی مانند انتشار، حیاتی است.
- یکی از محیط های مناسب برای تجزیه و تحلیل نفوذ اجتماعی میکرو بلاگ تویتر است که مطالعات متعددی با استفاده از آن برای ردیابی روندهای مختلف انجام شده است از جمله ردیابی اخبار (Lerman و Ghosh ، ۲۰۱۰؛ Petrovi'c و همکاران، ۲۰۱۰)، نظارت بر زلزله (Sakaki و همکاران، ۲۰۱۰)، احساسات (Barbosa و Feng ، ۲۰۱۰) و نظرات سیاسی (Tumasjan و همکاران، ۲۰۱۰؛ O'Connor و همکاران، ۲۰۱۰). به طور مشابه، توییت ها به موضوعات مرتبط با سلامت، مانند "من تب کرده ام من آنفلانزا گرفته ام" و موارد مشابه اشاره کرده است. این توییت ها نشان می دهد که کاربر دارای بیماری (آنفلانزا) علائم مرتبط (تب، و غیره) و درمان (تیلنول) است. خود گزارش دهی بهداشتی میلیون ها نفر از کاربران می تواند اطلاعات گسترده ی مربوط به سلامت جامعه به صورت روزانه فراهم کند (Paul و Dredze، ۲۰۱۲).

کارهای قبلی در این زمینه به طور خاص بر روی آنفلوانزا و ارزیابی نظارت آنفلوانزا (Lampos و Cristianini، 2010، Culotta، 2010b)، تجزیه و تحلیل توییت در مورد بیماری همه گیر H1N1 (Quincey و Kostkova، 2010)، و ترکیب بازار پیش بینی و توییت برای پیش بینی H1N1 (Ritterman و همکاران، ۲۰۰۹) متمرکز شده است. همه این نتایج از مدل های نظارت ساخته شده برای برنامه های کاربردی خاص بوجود می آیند (به عنوان مثال نظارت بر آنفلوانزا).

در اینجا چالش در ساخت یک الگوریتم است که بتواند به طور قابل اعتمادی میان این انواع "توییت" به طور خودکار و در زمان واقعی تفاوت ایجاد کند. Ni و همکاران (۲۰۰۷) توصیف یک راه حل با استفاده از روش یادگیرنده را برای طبقه بندی پست های وبلاگ به آموزنده و عاطفی ایجاد کرده اند (Kamel Boulos و همکاران، ۲۰۱۰).

همچنین مطالعات توییت نشان داده است که جمع آوری میلیون ها پیام می تواند بینش ارزشمندی را نسبت به مردم ارائه کند (Paul و Dredze، ۲۰۱۱). Barbosa و Feng (۲۰۱۰) توییت ها را متناسب با احساسات طبقه بندی کرده اند و اولین گام را در جهت سنجش افکار عمومی، مانند احساسات سیاسی برای پیگیری افکار عمومی سیاسی و پیش بینی نتایج انتخابات برداشته اند (Tumasjan و همکاران، ۲۰۱۰؛ اوکانر و همکارانش، ۲۰۱۰). Eisenstein و همکاران، (۲۰۱۰) تغییرات واژگانی را در سراسر مناطق جغرافیایی به طور مستقیم از توییت ها مورد مطالعه قرار دادند. نظارت بر گسترش اخبار (Lerman و Ghosh، ۲۰۱۰)، اولین اشاره به اخبار و رویدادها (Petrovi'c و همکاران، ۲۰۱۰) و شناسایی و نظارت بر زلزله (Sakaki و همکاران، ۲۰۱۰) از موارد دیگری در این زمینه است.

کاربران توییت اغلب در عرصه ی عمومی در پیامهایی مانند "من آنفلوانزا گرفته ام" و "آنفلوانزا تمام بدن مرا فراگرفته است" به بیان اطلاعات شخصی می پردازند. آگاهی از این که یک کاربر خاص آنفلوانزا گرفته است ممکن است جالب توجه نباشد، اما سنجش میلیون ها پیام می تواند باعث ردیابی نرخ آنفلوانزا در انگلستان و ایالات متحده آمریکا باشد (Lampos و Cristianini، ۲۰۱۰؛ Culotta، ۲۰۱۰).

علاوه بر این، توییت ها رویدادی جدا نیستند: آنها با زمان، مکان و زبان کاربران رخ می دهند. جمع آوری توییت های بیش از میلیون ها نفر از کاربران می تواند ابزارهای جدیدی برای تحقیقات بهداشت عمومی ارائه کند.

۴-۴. روند یابی

Eysenbach (۲۰۰۶؛ ۲۰۰۹) infodemiology را به عنوان "علم توزیع و عوامل تعیین کننده در اطلاعات در یک محیط الکترونیکی، به ویژه اینترنت، و یا در یک جمعیت، با هدف نهایی برای اطلاع

رسانی سلامت عمومی و سیاست عمومی" تعریف می کند. او در ادامه مفهوم مرتبط، infoveillance را به عنوان "علم مشابه که هدف اصلی آن نظارت است" تعریف می کند. Infodemiology و infoveillance در تجزیه و تحلیل خودکار و پیوسته به اطلاعات بدون ساختار، و متنی که در اینترنت در دسترس هستند، تکیه می کنند. این امر شامل تجزیه و تحلیل پرس و جو موتور جستجو (و یا آنچه Eysenbach "تقاضا" نامیده) و نیز از آنچه در حال انتشار بر روی وب است مانند وبلاگ ها، تویتر و غیره (و یا آنچه Eysenbach "عرضه" نامیده) می باشد.

بر اساس همین تعریف، محققان گوگل متوجه شدند که برخی از عبارات جستجو شاخص های خوبی را از فعالیت ویروس آنفلوآنزا نشان می دهند (Ginsberg و دیگران، ۲۰۰۹). Google Flu Trends (<http://www.google.org/flutrends>) با استفاده از داده های تجمیعی جستجو در گوگل به منظور برآورد میزان فعالیت ویروس آنفلوآنزا، تا دو هفته سریع تر از سیستم های رایج هشدار می دهد. چنین تشخیص زود هنگام فعالیت بیماری (در نزدیک به زمان واقعی)، هنگامی که با پاسخ سریع مناسب همراه شود، پتانسیل کاهش تاثیر هر دو آنفلوآنزای فصلی و همه گیر را داراست (Kamel Boulos و همکاران، ۲۰۱۰).

۴-۵. تجزیه و تحلیل پیش بینی شده اجتماعی (TPA)

رویدادهایی به صورت روزانه اتفاق می افتد که برای بهداشت، امنیت و رشد پایدار جامعه چالش برانگیز است و اغلب نداشتن آمادگی برای آن فاجعه بار است. این وقایع شامل تعامل فرآیندهای پیچیده مانند تغییرات آب و هوایی، بیماری های عفونی نوظهور، قابلیت اطمینان انرژی، تروریسم، گسترش سلاح های هسته ای، بلایای طبیعی و انسان ساخته شده و آسیب پذیری های جغرافیای سیاسی، اجتماعی و اقتصادی است (Kamel Boulos و همکاران، ۲۰۱۰).

TPA (Sanfilippo و دیگران، ۲۰۰۹) مدلی در حمایت از استدلال پیش بینی از طریق یک رویکرد چند بعدی در تجزیه و تحلیل استراتژیک و پاسخ است که هماهنگی تصمیم گیری تحلیلگران و سیاستگذاران را ممکن می سازد. این امر باعث افزایش آگاهی در میان کارشناسان موضوعی، تحلیل گران و تصمیم گیرندگان در مقابله با موقعیت های حیاتی می شود.

به عنوان مثال، مطالعات چند سطحی که شامل طیف گسترده ای از بیولوژیک، خانواده، جامعه، اجتماعی و فرهنگی، زیست محیطی، سیاست، و عوامل اقتصادی در سطح کلان برای پیشگیری و کاهش ظهور تهدیدات سلامت مانند چاقی در دوران کودکی و اعتیاد به مواد مخدر است، لازم است. درک مشترک از اثرات انسانی (به عنوان مثال ضایعات شیمیایی) و فرآیندهای طبیعی (به عنوان مثال تغییرات خورشیدی)

برای پیش بینی تاثیر گرم شدن کره زمین (Gore، ۲۰۰۶) مورد نیاز است. این بینش در رویکرد های شناختی و عاطفی انسان برای بهبود درک ما از تصمیمات اقتصادی و اثر در قیمت های بازار و تخصیص منابع لازم است (Mullainathan و Thaler، ۲۰۰۱). TPA پیشرفت های نظری و عملیاتی را از این بینش فراهم می کند.

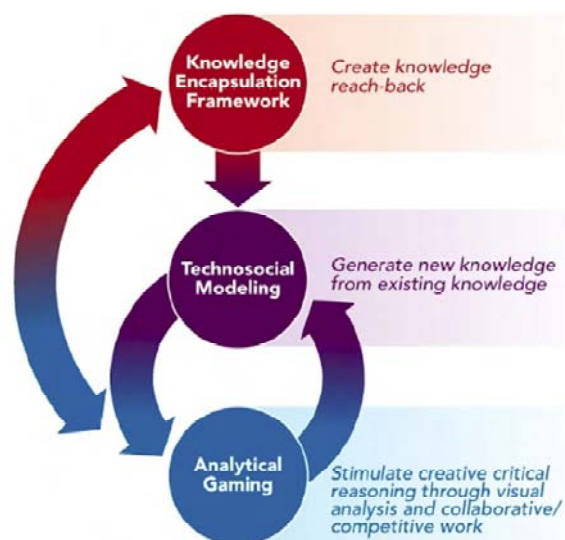
مغز انسان یک چارچوب اساسی برای حافظه و پیش بینی ایجاد می کند، که در آن تصمیم گیری شامل فرایندهای ذخیره سازی الگو، شناخت و طرح ریزی که ریشه در تجربه ما از جهان دارد و اجرا شدن آن توسط ادراک و خلاقیت ماست (Hawkins و دیگران، ۲۰۰۴). ویژگی هایی مانند توانایی تمرکز بر روی آنچه که تصور می شود بسیار مهم است "در لحظه" و بر اساس ظرفیت تصمیم گیری سریع از بینش و شهود باشد که قضاوت انسانی را منحصر به فرد می کند (Gigerenzer و Feelings، ۲۰۰۷؛ Gladwell، ۲۰۰۵).

TPA این شکاف ها را با استفاده از مدیریت دانش و مدل سازی برای اطلاع رسانی کافی به فرد و تحلیل مناسب پر می کند.

به طور خاص، TPA تحلیلگران و سیاستگذاران را در ارائه تجزیه و تحلیل پیشگیرانه و پاسخ در گرفتن تصمیمات واقعی کمک می کند. این امر از طریق مجموعه ای از قابلیت های زیر ایجاد می گردد:

(الف) ارائه دانش چند بعدی برای اطلاع کافی در تجزیه و تحلیل و پاسخ در هنگام تصمیم گیری
(ب) پشتیبانی از تخصص تحلیلگر و قانونگذاران با سناریوهای شبیه سازی شده توسط مدل های محاسباتی از ادغام عوامل انسانی و فیزیکی

(ج) تعامل با تحلیلگران و سیاستگذاران در محیطی بازی مانند که باعث تحریک استدلال انتقادی خلاقانه از طریق تجزیه و تحلیل بصری مشترک یا رقابتی می شود.



شکل ۲-۴. روابط بین مولفه های TPA (Kamel Boulos و همکاران، ۲۰۱۰)

همانطور که در شکل ۴-۲ نشان داده شده TPA یک محیط بصری تحلیلی و همکاری با دسترسی کاربر پسند به مدل های پیش بینی شده در زمینه های امنیتی، انرژی و محیط زیست است که توسط فرآیندهای مدیریت دانش ایجاد می شود.

چنین محیطی باعث یکپارچه سازی مدل ها از طریق الگوریتم های نوآورانه می شود که ترکیبی از الگوها و مدل سازی های مختلف مانند شبکه های Bayesian و پویایی سیستم را قادر می سازد. مدل سازی تجمیع داده ها با یگدیگر از KEF استفاده می کند. که باعث تسهیل استخراج و تأیید دانش مربوطه از جریان داده های ناهمگن، از جمله در رسانه های اجتماعی است.

TPA، یک رویکرد ایده آل علمی برای انفورماتیک سلامت عمومی است که به چالش هایی مانند بیماری های عفونی نوظهور اشاره می کند. مدل های اپیدمی کنونی تاثیرات حمل و نقل، صفات فردی، مدارس، دسترسی به مراقبت های بهداشتی را مشخص می کنند. TPA می تواند این الگوها و مدل سازی موجود را از طریق ادغام مدل های اجتماعی و فرهنگی گسترش دهد و به ارائه خصوصیات دقیق تری که منعکس کننده عواملی مانند نظم اجتماعی، فرد و جامعه مبتنی بر دانش سلامت و تغییر رفتار با توجه به ارتباطات سلامت عمومی در رسانه های جمعی و فن آوری های وب نو ظهور است، پردازد.

یک کاربرد خاص به روش TPA انفورماتیک سلامت عمومی و نظارت بر آنفلوآنزای فصلی از طریق وب سایت و رسانه های اجتماعی است که می تواند از منابع اطلاعاتی مختلف مانند وبلاگ ها و یا سیستم های تجمیع اطلاعات وبلاگ ها و محیط های اجتماعی مانند spinn3r.com استفاده کند.

۴-۶. مدل موضوعی بیماری (ATAM)

مدل های موضوع احتمالی مانند تخصیص نهفته دیریکله (LDA) (Blei و همکاران، ۲۰۰۳)، کلمات را با موضوعاتی مرتبط می کنند. اسنادی که در موضوعات مختلفی منتشر می شوند اغلب تشکیل یک مجموعه کلمات معنایی منسجم می دهند. ATAM مدلی است که چگونگی بیان بیماری ها و بیماری های خود را در tweets نشان می دهد که بر اساس موضوعات عمل می کند. این مدل فرض می کند که هر tweet مرتبط با سلامت نشان دهنده یک بیماری نهفته مانند آنفلوآنزا، آلرژی، یا سرطان است.

علاوه بر بیماری، از آنجایی که کاربران در tweet ها در مورد موضوعات بهداشتی و اقدامات نامربوط به درمان بحث می کنند مثلاً "امروز بیمارم پس فقط بازی های ویدئویی"، ATAM شامل مولفه های رایج و طبیعی دیگری نیز هست که باید شناسایی شود که هر چند ساختار مدل متفاوت است اما الگوریتم نمونه گیری گیبس می تواند در بهبود پارامترهای ATAM استفاده شود.

۴-۵. خلاصه فصل

استفاده از داده ها در اینترنت برای مدل سازی و تجزیه و تحلیل رفتارهای بهداشت عمومی، روشی کارآمد در مقالات گذشته بوده است. Google Flu Trends پیش بینی های متنوعی از ابتلا به بیماری آنفلوآنزای بر اساس پرس و جو آنلاین فراهم می کند. Paul و Dredze (۲۰۱۱) یک مدل با موضوع بیماری خاص را بر اساس پست های تویتر به منظور مدل رفتار انواع بیماری در اهمیت در بهداشت عمومی توسعه داده اند. از طریق مدل سازی زبان از پست های تویتر، Collier و همکاران (۲۰۱۱) شواهدی از ارتباط بین سیگنال های رسانه های اجتماعی و داده های واقعی آنفلوآنزا سنجش کرده اند. Sadelik و همکاران (۲۰۱۲) مدل آماری توسعه داده اند که پیش بینی بیماری های عفونی (مانند آنفلوآنزا) و گسترش آن را در افراد بر اساس موارد دارای برچسب جغرافیایی در تویتر سنجش می کنند (Choudhury و همکاران، ۲۰۱۳). در حالی که این تحقیقات به بررسی طیف وسیعی از چالش ها در سراسر بهداشت عمومی می پردازند، پژوهش در زمینه بهره برداری از رسانه های اجتماعی برای درک اختلالات سلامت رفتاری هنوز در مراحل ابتدایی است. Park و همکاران (۲۰۱۲) شواهد اولیه ای را در اینکه مردم پست های در مورد افسردگی و حتی درمان آن در تویتر منتشر می کنند، بررسی کرده اند. در این فصل به بررسی انواع مدل های استدلالی برای نظارت و پیشگیری در حوزه سلامت عمومی پرداخته شد و مزایا و معایب هر مدل نیز بیان گردید.

فصل پنجم - نتیجه گیری و جمع بندی

۵-۱. مقدمه

اصطلاح Web2.0 اولین بار توسط Tim O'Reilly در سال ۲۰۰۵ برای اشاره به نسل جدید از فن آوری های وب، ابداع شد (Meier و همکاران، ۲۰۰۷). این فن آوری ها شامل سایت های شبکه های اجتماعی که مشوق همکاری و بحث و گفتگو میان کاربران است، سایتهایی که توسط دانش جمعی ایجاد شده است مانند ویکی پدیا، و mashup ها، که اجازه می دهند برنامه های کاربردی وب به راحتی برای ایجاد ارزش بیشتر ترکیب شوند، می شوند (Chun و همکاران، ۲۰۱۲). برنامه های کاربردی همراه، با افزایش در دسترس بودن دستگاه های تلفن همراه که اینترنت دارند (مانند نوت بوک و آی فون اپل)، باعث شده است که اینترنت در حال حاضر واقعا بخشی جدایی ناپذیر از زندگی روزمره باشد. وب اجتماعی رسانه ای خصوصا پویا است که پالس های جامعه را در زمان واقعی سنجش میکند. وبلاگ ها، میکرو وبلاگ مانند توییتر و سایت های شبکه های اجتماعی مانند فیس بوک، مردم را قادر به انتشار داستان های شخصی خود، نظرات، نظرات در مورد یک محصول و موارد دیگر در زمان حال می کند. برنامه هایی از web2.0 که در زمینه مراقبت های بهداشتی است فن آوری Health2.0 یا Medicine2.0 نامیده می شوند.

متن کاوی جستجوها در گوگل نشان داده است که مطالب ارسالی در مورد آنفلونزا با روند بیماری آنفلونزا در دنیای واقعی یکسان است و مطالعات متعددی با استفاده از توییتر برای ردیابی روندهای مختلف انجام شده اند.

همچنین در رابطه با سلامت به طور خاص، سیستم های ارتباطی در رسانه های اجتماعی و وب برای به دست آوردن اطلاعات مفید بهداشتی و سلامت عمومی خصوصا با توجه به اینکه کاربران نهایی می توانند از طریق دستگاه های تلفن همراه و وسایل مشابه به اینترنت به طور مداوم متصل شوند و اطلاعاتی از

سلامت و بهداشت خود منتشر کنند، نقش مهمی دارند. همچنین سکوهایی مانند توییتر برای محققان فرصت کافی برای تحلیل پدیده های اجتماعی فراهم آورده است. استفاده از این ظرفیت جهت نظارت عمومی بر بهداشت و سلامت با توجه به اینکه در ایران نفوذ اینترنت در حال افزایش است، اهمیت بالایی پیدا کرده است. البته چالش هایی در این زمینه وجود دارد از جمله اعتبار سنجی اطلاعات و همچنین مشکلاتی که در رابطه با زبان فارسی در داده کاوی وب وجود دارد. زبان فارسی، در تلاقی با جهان الکترونیکی، بخصوص از بعد رسم الخط، دارای مشکلاتی است که کاوش در محتویات آن را دچار کم کیفیتی می نماید.

۲-۵. مروری بر فصل های گذشته

در این سمینار به بررسی چگونگی اتصال داده های پراکنده مرتبط با سلامت از جوامع مختلف وب و ارائه دانش یکپارچه اطلاعات سلامت (به طور خاص از داده های رسانه های اجتماعی مبتنی بر جوامع بیماران، سایت های مراقبت درمانی با محتوای تخصصی)، به منظور نظارت بر بهداشت عمومی و پیش بینی اتفاقاتی مثل شیوع بیماری ها از طریق روش های همچون داده کاوی متن و ساختار وب در رسانه های اجتماعی (WSM) (Corley و همکاران، ۲۰۱۰) که می تواند یک منبع نظارتی برای بیماری فراهم کند و می تواند جوامع آنلاین برای ارتباطات هدفمند بهداشت عمومی را (PHC) شناسایی کند، سیستم بازیابی اطلاعات وب (WebIRS) که به متخصصان مراقبت های بهداشتی برای به دست آوردن اطلاعات به روز دانش پزشکی و اطلاعات از طریق شبکه جهانی وب (WWW) کمک می کند (Ting و همکاران، ۲۰۱۳) و تجزیه و تحلیل پیش بینی اجتماعی (TPA) پرداخته شد.

در فصل اول راجع به تحولات وب و اهمیت موضوع شناسایی بیماری از طریق وب اشاره شده و در ادامه در فصل دوم منابع مختلف اطلاعات سلامت بر روی وب بررسی شده اند که این منابع شامل: شبکه های اجتماعی، وبلاگ ها، اخبار، روند های جستجو و میکرو بلاگ ها می باشند.

در فصل سوم به چالش های استخراج اطلاعات از این منابع و مشکلات زبان فارسی پرداخته شده است و به روند اجرای فرایند سیستم های خزنده (Crawler) اشاره شده است.

در فصل چهارم به روند های پیش بینی بر اساس اطلاعات استخراج شده از منابع مختلف وب پرداخته شده و روش هایی همچون داده کاوی متن و ساختار وب در رسانه های اجتماعی (WSM)، سیستم بازیابی اطلاعات وب (WebIRS) و تجزیه و تحلیل پیش بینی اجتماعی (TPA) بیان شده است.

۳-۵. مروری بر تحلیل و نقد ادبیات موضوع

همانطور که اشاره شد تحقیقات مختلفی در زمینه استفاده از منابع مختلف وب و روش های مختلف استخراج اطلاعات از وب جهت شناسایی بیماری ها و شیوع بیماری ها انجام شده است. از مهم ترین این منابع وب سایت توییتر و سایت هایی همچون فیسبوک بوده است. در بعضی از تحقیقات رسانه های اجتماعی به عنوان یک ابزار جدید برای اندازه گیری سلامت روانی و نظارت معرفی شده اند و نشان داده اند که همبستگی مناسب بین اطلاعات بدست آمده با اطلاعات گرفته شده از منابع رسمی وجود دارد.

در فرایند استخراج اطلاعات، به ابزار مناسبی جهت نمونه گیری و نحوه نمونه گیری از داده های منابع مختلف وب خصوصاً توییتر پرداخته شده که نشان دهنده اهمیت این منبع برای استفاده در پیش بینی ها بر اساس محیط اجتماعی وب است.

همچنین روش های تحلیل مختلفی برای اطلاعات جمع آوری شده استفاده شده است از جمله روش تجزیه و تحلیل پیش بینی اجتماعی (TPA)، داده کاوی متن و ساختار وب و رسانه های اجتماعی (WSM) و سیستم بازیابی اطلاعات وب (WebIRS).

لازم به ذکر است تمامی این تحقیقات در محدوده یک دهه گذشته و پس از انقلاب Web2.0 که منجر به تولید محتوا در محیط گسترده وب توسط کاربران شد، بوده است.

۴-۵. محدودیت ها و نکات قابل توجه

مهم ترین محدودیت هایی که در مرور ادبیات به نظر رسید شامل موارد زیر است:

۱. عدم وجود یک سیستم فارسی برای جمع آوری اطلاعات همچون سایت spinn3r.com که بتوان بر اساس آن به اطلاعات جامعی در رابطه با وبلاگ ها و شبکه های خبری در حوزه سلامت رسید.
۲. محدودیت در دسترسی به شبکه های اجتماعی چون فیسبوک و میکرو بلاگ هایی چون توییتر در ایران که باعث کاهش نسبت کاربران به نسبت کل جمعیت یک کشور و خطای تشخیص با استفاده از چنین محیط هایی می شود.
۳. عدم وجود یک بانک کلمات کلیدی مناسب در زمینه پزشکی به زبان فارسی (در مقالات یافت شده اشاره نشده بود) که در تحلیل های متن کاوی صفحات وب بسیار موثر است.
۴. عدم وجود یک تالار گفتمان یا انجمن بحث در حوزه های پزشکی که دارای تعداد کاربران متناسبی با جمعیت کشور باشد که بتوان از طریق داده های آن به تحلیلی در نظارت بر سلامت عمومی رسید.

۵-۵. زمینه ها و موضوعات پژوهشی برگرفته از مطالعه

زمینه های پیشنهادی با توجه به محدودیت های گفته شده شامل موارد زیر است:

۱. راه اندازی سیستم جمع آوری اطلاعات وبلاگ های فارسی زبان
۲. استفاده از روندیابی گوگل و api گوگل جهت کسب اطلاعات نظارتی سلامت
۳. راه اندازی سیستم نقشه سلامت ایران بر مبنای اطلاعات کسب شده از سیستم های خبری
۴. ایجاد پایگاه کلمات کلیدی در حوزه سلامت وب به زبان فارسی

۵-۶. طرح و بیان مسئله تحقیق پایان نامه

آنچه به عنوان مساله پایان نامه می تواند قرار گیرد شامل موارد زیر است:

۱. آیا می توان از طریق اطلاعات سلامت محتوای فارسی وب سایت ها در سنجش سلامت عمومی در ایران استفاده کرد؟
۲. امکان استفاده از api گوگل جهت روندیابی و نظارت بر بیماری ها در ایران چگونه است؟
۳. چگونه می توان از طریق منابع مختلف اطلاعاتی در وب، نقشه جامع سلامت در ایران را ایجاد نمود؟

۵-۷. زمینه های کاربردی پیشنهادی و میزان دسترسی به داده های تحقیق

با توجه به زمینه های پژوهشی مطرح شده زمینه های کاربردی پژوهش در حوزه نظارت بر سلامت عمومی و تهیه نقشه جامع سلامت در ایران می باشد.

۵-۷-۱. استراتژی تحقیق

مبنای پژوهش، منابع اطلاعاتی در وب در حوزه سلامت که به صورت متون غیر ساخت یافته و یا نیمه ساخت یافته است می باشد و برای اینکار از روش های متن کاوی و مدل های ارائه شده در تحقیق استفاده می گردد.

۵-۷-۲. روش گردآوری داده ها و جامعه آماری

برای جمع آوردی داده ها با توجه به محدودیت هایی که ذکر شد و عدم دسترسی کاربران در ایران به محیط هایی همچون facebook و يا tweeter، برای جمع آوردی داده های خام از تالارگفتمان ها و وبلاگ های فارسی زبان و استفاده از api گوگل جهت روندیابی کلمات کلیدی حوزه سلامت به زبان فارسی استفاده می شود.

۵-۸. خلاصه فصل

در این فصل ابتدا به جمعبندی مطالب مطرح شده در فصول گذشته پرداخته شده و سپس با مروری بر ادبیات موضوع، پیشنهاداتی به همراه محدودیت هایی برای پژوهش بیان شده است و در نهایت با بررسی مساله پایان نامه و زمینه های کاربردی طرح و بررسی داده های پژوهش، فصل به اتمام رسیده است.

فهرست منابع

منابع فارسی:

۱. خاصه، ع.ا.، ۱۳۸۸. داده کاوی، متن کاوی، و وب کاوی: تعاریف و کاربردها.
۲. شهیدی، س.م.، صدیقی، م.، زمانی فر، ک.، ۱۳۸۴. روشی برای رفع چالش های محتوا کاوی وب های فارسی زبان. مرکز اطلاعات و مدارک علمی ایران، ۲۴، ۱۸-۳۰.

منابع لاتین:

3. Anagnostopoulos, A., Kumar, R., Mahdian, M., 2008. Influence and correlation in social networks. In: KDD '08. pp. 7–15.
4. Andreassen, H.K., Bujnowska-Fedak, M.M., Chronaki, C.E., Dumitru, R.C., 2007. European citizens' use of E-health services: a study of seven countries. BMC Public Health.
5. Antonio do Prado, H., Ferneda, E., 2008. Emerging technologies of text mining: techniques and applications. Information Science Reference, U.S.A.
6. Aron, C., McCallum, A., 2004. Confidence estimation for information extraction. In: Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics.
7. Arotaritei, D., Mitra, S., 2004. Web mining: a survey in the fuzzy framework. Fuzzy Sets and Systems 148, 5–19.
8. Baker, L., Wagner, T.H., Singer, S., Bundorf, M.K., 2003. Use of the Internet and e-mail for health care information: results from a national survey. JAMA.
9. Bakshy, E., Karrer, B., Adamic, L.A., 2009. Social influence and the diffusion of user-created content. In: EC '09. New York, NY, USA, pp. 325–334.
10. Barbosa, L., Feng, J., 2010. Robust sentiment detection on twitter from biased and noisy data. In: COLING.
11. Berry, M.W., 2004. Survey of text mining: clustering, classification and retrieval. Springer.
12. Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3.

13. Boyd, D., Golder, S., Lotan, G., 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In: 43rd Hawaii International Conference on Systems Sciences.
14. Breast cancer discussion board [WWW Document], 2011. . WebMD.com. URL <http://exchanges.webmd.com/breast-cancer-exchange>
15. Brownstein, J.S., Freifeld, C.C., Madoff, L.C., 2009. Digital Disease Detection — Harnessing the Web for Public Health Surveillance. *The NEW ENGLAND JOURNAL of MEDICINE* 2153–2157.
16. Callison-Burch, C., Dredze, M., 2010. Creating speech and language data with amazon’s mechanical turk. In: Workshop on Creating Speech and Language Data With Mechanical Turk at NAACL-HLT.
17. Carneiro, H., Mylonakis, E., 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin Infect Dis* 49, 1327–1328.
18. Choudhury, M.D., Counts, S., Horvitz, E., 2013. Social Media as a Measurement Tool of Depression in Populations.
19. Choudhury, M.D., Lin, Y.-R., Sundaram, H., Candan, K.S., Xie, L., Kelliher, A., 2010. How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media? In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp. 34–41.
20. Chun, S.A., MacKellar, B., 2012. Social health data integration using semantic Web. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing. pp. 392–397.
21. Chunara, R., Andrews, J.R., Brownstein, J.S., 2012. Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *The American Society of Tropical Medicine and Hygiene* 86, 39–45.
22. Collier, N., Son, N., Nguyen, N., 2011. OMG U got flu? Analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics*.
23. Corley, C.D., Cook, D.J., Mikler, A.R., Singh, K.P., 2010. Text and Structural Data Mining of Influenza Mentions in Web and Social Media. *International Journal of Environmental Research and Public Health* 596–615.
24. Craan, F., Oleske, D.M., 2002. Medical information and the Internet: do you know what you are getting? *J. Med. Syst* 26, 511–518.
25. Culotta, A., 2010. Detecting influenza epidemics by analyzing twitter messages. *arXiv*.
26. Eisenstein, J., O’Connor, B., Smith, N.A., Xing, E.P., 2010. A latent variable model for geographic lexical variation. In: Empirical Natural Language Processing Conference.

- 27.Elhadad, N., Kan, M.Y., Klavans, J.L., McKeown, K.R., 2005. Customization in a unified framework for summarizing medical literature. *Artif. Intell. Med.* 33, 179–198.
- 28.Etter, P. von, Huttunen, S., Vihavainen, A., Vuorinen, M., Yangarber, R., 2010. Assessment of utility in web mining for the domain of public health. In: *Proceedings of the Louhi '10 NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*. pp. 29–37.
- 29.Eysenbach, G., 2005. Annual Symposium. In: *Proceedings of the AMIA Annual Symposium*. Washington, DC, USA, pp. 244–248.
- 30.Eysenbach, G., 2008. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *J Med Internet Res*.
- 31.Eysenbach, G., 2009. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *J Med Internet Res* 11.
- 32.Eysenbach, G., Kohler, C., 2003. Is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the internet. *AMIA Annu Symp Proc*.
- 33.Fernandez-Luque, L., Karlsen, R., Bonander, J., 2011. Review of extracting information from the social web for health personalization. *Journal of Medical Internet Research* 13.
- 34.Fox, S., 2006. Online Health Search 2006 [WWW Document]. Pew Internet & American Life Project. URL <http://www.pewinternet.org/Reports/2006/Online-Health-Search-2006.aspx?r=1>
- 35.Fox, S., 2011. The Social Life of Health Information. Pew Internet and American Family Project.
- 36.Fox, S., Jones, S., 2009. The social life of health information [WWW Document]. Pew Internet & American Life Project. URL http://www.pewinternet.org/~media/Files/Reports/2009/PIP_Health_2009.pdf
- 37.Frost, J., Massagli, M., 2008. Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data. *J Med Internet Res*.
- 38.G. B., B., Sethi, A., 2001. Matching records in a national medical patient index. *Commun. ACM* 44, 83–88.
- 39.Gaster, B., Knight, C.L., DeWitt, D.E., Sheffield, J.V.L., Assefi, N.P., Buchwald, D., 2003. Physicians' use of and attitudes toward electronic mail for patient communication. *J Gen Intern Med*.

40. Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 1012–1014.
41. Greene, J., Choudhry, N., Kilabuk, E., Shrank, W., 2010. Online social networking by patients with diabetes: A qualitative evaluation of communication with facebook. *Journal of General Internal Medicine*.
42. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A., 2004. Information diffusion through blogspace. In: *WWW '04*. New York, NY, USA, pp. 491–501.
43. Han, J., Kamber, M., 2006. *Data mining: concept and techniques*. Morgan Kaufmann, San Francisco.
44. Hartley, D., Nelson, N., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., 2010. The landscape of international event-based biosurveillance. *Emerging Health Threats Journal*.
45. Hesse, B.W., Nelson, D.E., Kreps, G.L., Croyle, R.T., 2005. Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. *Arch Intern Med* 165, 2618–2624.
46. Hirschman, L., 1998. Language understanding evaluations: Lessons learned from muc and atis. In: *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*. Granada, Spain, pp. 117–122.
47. Hughes, B., Joshi, J., Wareham, J., 2008. Health 2.0 and Medicine 2.0: Tensions and controversies in the field. *Journal of Medical Internet Research* 10.
48. Hulth, A., Rydevik, G., Linde, A., Montgomery, J., 2009. Web Queries as a Source for Syndromic Surveillance. *PLoS ONE* 4.
49. Java, A., Song, X., Finin, T., Tseng, B., 2010. Why we twitter: An analysis of a microblogging community. In: *Advances in Web Mining and Web Usage Analysis, Volume Lecture Notes in Artificial Intelligence*.
50. Kamel Boulos, M.N., Sanfilippo, A.P., Corley, C.D., Wheeler, S., 2010. Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. *computer methods and programs in biomedicine* 100, 16–23.
51. Kamel Boulos, M.N., Scotch, M., Cheung, K.-H., Burden, D., 2008. Web GIS in practice VI: a demo “playlist” of geo-mashups for public health neogeographers. *International Journal of Health Geographics* 38.
52. Kamel Boulos, M.N., Wheeler, S., 2007. The emerging Web 2.0 social software: an enabling suite of sociable technologies in health and healthcare education. *Health Information and Libraries Journal* 24, 2–23.

53. Kayumba, K., Adeline, K., Baptiste, K.J., Anita, A., 2013. TRACnet: A National Phone-based and Web-based Tool for the Timely Integrated Disease Surveillance and Response in Rwanda. *Online Journal of Public Health Informatics*.
54. Kempe, D., Kleinberg, J., Tardos, E., 2003. Maximizing the spread of influence through a social network. In: *KDD '03*. New York, NY, USA, pp. 137–146.
55. King, D., Ramirez-Cano, D., Greaves, F., Vlaev, I., Beales, S., Darzi, A., 2013. Twitter and the health reforms in the English National Health Service. *Health Policy* 110, 291–297.
56. Ko, J., Lu, C., Mani, S.B., Stankovic, J., Terzis, A., Welsh, M., 2010. Wireless sensor networks for healthcare. In: *Proc. IEEE*. pp. 1947–1960.
57. Kossinets, G., Kleinberg, J., Watts, D.J., 2008. The structure of information pathways in a social communication network. In: *KDD '08*. pp. 435–443.
58. Ku, Y., Chiu, C., Liou, B.H., 2008. Applying text mining to assist people who inquire HIV/AIDS information from Internet. In: *Proceedings ISI 2008 Workshops*. pp. 440–448.
59. Kumar, R., Novak, J., Tomkins, A., 2006. Structure and evolution of online social networks. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. Philadelphia, PA.
60. Kummervold, P., Chronaki, C., Lausen, B., Prokosch, H., Rasmussen, J., Santana, S., Staniszewski, A., Wangberg, S., 2008. eHealth trends in Europe 2005-2007: a population-based survey. *J Med Internet Res*.
61. Lampos, V., Cristianini, N., 2010. Tracking the flu pandemic by monitoring the social web. In: *IAPR 2nd Workshop on Cognitive Information Processing*.
62. Lerman, K., Ghosh, R., 2010. Information contagion: an empirical study of the spread of news on digg and twitter social networks. In: *CoRR*.
63. Liben-Nowell, D., Kleinberg, J., 2008. Tracing information flow on a global scale using internet chain-letter data. In: *Proc. National Academy of Sciences* 105. pp. 4633–4638.
64. Liu, Z., Chu, W.W., 2007. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Inf. Retr.* 10, 173–202.
65. Logan, R.A., 2004. Evaluating consumer informatics: learning from health campaign research. *Stud Health Technol Inform* 1147–1151.
66. Lu, W.H., Lin, R.S., Chan, Y.C., Chen, K.H., 2008. Using Web resources to construct multilingual medical thesaurus for crosslanguage medical information retrieval. *Decis. Support. Syst.* 45, 585–595.

67. Luque, L.F., Karlsen, R., Bonander, J., 2011. Review of Extracting Information From the Social Web for Health Personalization. *JOURNAL OF MEDICAL INTERNET RESEARCH* 13.
68. MacLean, D., Seltzer, M., 2011. Mining the Web for medical hypotheses: A proof-of-concept system. In: *Proceedings of the 2011 International Conference on Health Informatics (HEALTHINF'11)*. Rome, Italy.
69. Manning, C.D., Raghavan, P., Schütze, H., 2008. *Introduction to information retrieval*. Cambridge University Press, U.K.
70. Marban, O., Et al., 2009. Toward data mining engineering: a software engineering approach. *Information Systems* 34, 87–107.
71. Markov, Z., Russell, I., 2009. An introduction to the WEKA data mining system [WWW Document]. URL <http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf>
72. Meier, A., Lyons, E.J., Frydman, G., Forlenza, M., Rimer, B.K., 2007. How cancer survivors provide support on cancer-related internet mailing lists. *Journal of Medical Internet Research* 9.
73. Mostafa, J., Lam, W., 2000. Automatic classification using supervised learning in a medical document filtering application. *Inf. Process. Manag.* 36, 415–444.
74. Newman, M.E.J., 2002. Spread of epidemic disease on networks. *Physical Review E* 66.
75. Nicholson, S., 2003. The Bibliomining Process: Data Warehousing and Data Mining for Library Decision making. *Information Technology and Libraries* 22, 146–151.
76. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A., 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In: *ICWSM*.
77. Ochneva, A., 2013. Web-application for gathering, analyzing, and processing health information about allergy data.
78. Onnela, J.-P., Arbesman, S., Barabási, A.-L., Christakis, N.A., 2010. Geographic constraints on social network groups. *arXiv*.
79. Park, M., Cha, C., Cha, M., 2012. Depressive Moods of Users Captured in Twitter. In: *Proc. ACM SIGKDD Workshop on Healthcare Informatics*.
80. Parr, B., 2011. 40% of all tweets come from mobile [WWW Document]. URL <http://mashable.com/2011/01/07/40-of-all-tweets-come-from-mobile/>
81. Patil, S.B., Kumaraswamy, Y.S., 2009. Intelligent and effective heart attack prediction system using data mining and artificial neural network. *Eur. J. Sci. Res.* 31, 642–656.
82. Paul, M., Girju, R., 2010. A two-dimensional topic aspect model for discovering mult faceted topics. In: *AAAI*.

83. Paul, M.J., Dredze, M., 2011. You Are What You Tweet: Analyzing Twitter for Public Health. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. pp. 265–272.
84. Paul, M.J., Dredze, M., 2012. A Model for Mining Public Health Topics from Twitter. HEALTH.
85. Perreault, M., Ruths, D., 2011. The effect of mobile platforms on Twitter content generation. In: Proceedings of the Fifth Intl. Conference on Weblogs and Social Media. Barcelona.
86. Petrović, S., Osborne, M., Lavrenko, V., 2010. Streaming first story detection with application to twitter. In: NAACL.
87. PHILIP, J.J., 2012. PROMINENT STREAKS DISCOVERY ON BLOG ARTICLES.
88. Polgreen, P., Chen, Y., Pennock, D., Nelson, F., 2008. Using internet searches for influenza surveillance. Clin. Infect. Dis. 47, 1443–1448.
89. PubMed - NCBI [WWW Document], 2011. . PubMed.gov. URL <http://www.ncbi.nlm.nih.gov/pubmed/>
90. Quincey, E., Kostkova, P., 2010. Early warning and outbreak detection using social networking websites: The potential of twitter. In Electronic Healthcare. Springer.
91. Ritter, A., Cherry, C., Dolan, B., 2010. Unsupervised Modeling of Twitter Conversations. In: NAACL.
92. Rogers, S.N., Rozek, A., Aleyaasin, N., Promod, P., Lowe, D., 2012. Internet use among head and neck cancer survivors in the North West of England. Oral Maxillofac. Surg. 50, 208–214.
93. Sadilek, A., Kautz, H., Silenzio, V., 2011. Modeling Spread of Disease from Social Interactions. In: Proc. ICSWM '11.
94. Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In: WWW. New York, NY, USA.
95. Seifter, A., Schwarzwald, A., Geis, K., Aucott, J., 2010. The utility of “google trends” for epidemiological research: Lyme disease as an example. Geospatial Health 4, 135–137.
96. Skeels, M., Grudin, J., 2009. When social networks cross boundaries: a case study of workplace use of Facebook and LinkedIn. In: Proc GROUP. pp. 95–104.
97. Song, M., Wu, Y.F., 2009. Handbook of research on text and web mining technologies. Information Science Reference, U.S.A.
98. Song, X., Tseng, B.L., Lin, C.-Y., Sun, M.-T., 2006. Personalized recommendation driven by information flow. In: SIGIR '06. pp. 509–516.
99. Steele, R., 2011. Social Media, Mobile Devices and Sensors: Categorizing New Techniques for Health Communication. Sensing Technology (ICST), 2011 Fifth International Conference on 187 – 192.

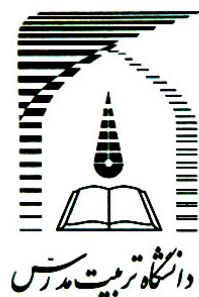
100. Sun, E., Rosenn, I., Marlow, C., Lento, T., 2009. Gesundheit! modeling contagion through facebook news feed. In: ICWSM '09.
101. T., H., Bell, A.A., 2001. Measuring and managing knowledge. McGraw Hill, Irwin.
102. Tian, H., Brimmer, D.J., Lin, J.-M.S., Tumpey, A.J., Reeves, W.C., 2009. Web Usage Data as a Means of Evaluating Public Health Messaging and Outreach. JOURNAL OF MEDICAL INTERNET RESEARCH 11.
103. Ting, S.L., See-To, E.W.K., Tse, Y.K., 2013. Web Information Retrieval for Health Professionals. J Med Syst 37.
104. Ting, S.L., Shum, C.C., Kwok, S.K., Tsang, A.H.C., Lee, W.B., 2009. Data mining in biomedicine: current applications and further directions for research. J. Softw. Eng. Appl. 2, 150–159.
105. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M., 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In: ICWSM.
106. Vakali, A., Pallis, G., 2007. Web data management practices: emerging techniques and technologies. Idea Group Publishing, U.S.A.
107. Vance, K., Howe, W., 2009. Social internet sites as a source of public health information. Dermatologic Clinics 27, 133–136.
108. Velasquez, J.D., Palade, V., 2008. Adaptive web sites: a knowledge extraction from web data approach. Ios Press, Netherlands.
109. Watts, D.J., Dodds, P.S., 2007. Influentials, networks, and public opinion formation. Journal of Consumer Research 34, 441–458.
110. Wheeler, S., Kamel Boulos, M.N., 2007. Mashing, burning, mixing and the destructive creativity of Web 2.0: applications for medical education. RECIIS—Electronic Journal of Communication, Information and Innovation in Health 27–33.
111. Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., Bradley, R., Heywood, J., 2010. Sharing Health Data for Better Outcomes on PatientsLikeMe. Journal of Medical Internet Research 12.
112. Williams, S.A., Terras, M.M., Warwick, C., 2013. What do people study when they study Twitter? Classifying Twitter related academic papers. Journal of Documentation 69, 384–410.
113. Yih, W., Teates, K., Abrams, A., Kleinman, K., Kulldorff, M., Pinner, R., Harmon, R., Wang, S., Platt, R., Montgomery, J., 2009. Telephone triage service data for detection of influenza-like illness. PLoS ONE 4.
114. Zachary, W., 1977. An information flow model for conflict and fission in small groups. . Journal of Anthropological Research 33, 452–473.

پیوست ۱- فهرست اصطلاحات فارسی به انگلیسی

Wiki	آنچه من می دانم این است
News	اخبار
IE	استخراج کننده اطلاعات
Blog	بلاگ
Disease	بیماری
ILI	بیماری های شبه آنفلوانزا
LDA	تخصیص نهفته دیریکله
RDF	چارچوب توضیح منابع
KEF	چارچوب دانش
RSS	خلاصه غنی از سایت
Data Mining	داده کاوی
PHC	روابط سلامت عمومی
MedISys	سیستم اطلاعات پزشکی
WebIRS	سیستم بازیابی اطلاعات وب
PULS	سیستم فهم و آموزش الگو محور
UMLS	سیستم یکپارچه زبان پزشکی
Social Network	شبکه اجتماعی
WWW	شبکه گسترده جهانی
Detect	شناسایی
ATAM	مدل موضوعی بیماری
CDC	مرکز کنترل و پیشگیری از بیماری
Web	وب
WSM	وب و محیط اجتماعی

پیوست ۲- فهرست اصطلاحات انگلیسی به فارسی

ATAM	مدل موضوعی بیماری
Blog	بلاگ
CDC	مرکز کنترل و پیشگیری از بیماری
Data Mining	داده کاوی
Detect	شناسایی
Disease	بیماری
IE	استخراج کننده اطلاعات
ILI	بیماری های شبه آنفلوانزا
KEF	چارچوب دانش
LDA	تخصیص نهفته دیریکله
MedISys	سیستم اطلاعات پزشکی
News	اخبار
PHC	روابط سلامت عمومی
PULS	سیستم فهم و آموزش الگو محور
RDF	چارچوب توضیح منابع
RSS	خلاصه غنی از سایت
Social Network	شبکه اجتماعی
UMLS	سیستم یکپارچه زبان پزشکی
Web	وب
WebIRS	سیستم بازیابی اطلاعات وب
Wiki	آنچه من می دانم این است
WSM	وب و محیط اجتماعی
WWW	شبکه گسترده جهانی



Tarbiat Modares University

**Faculty of Engineering & Technology
Industrial Engineering Department**

MSc Seminar

Disease Detection by Web

**Author
Said Ali Lajevardy**

**Master
Dr. M. Karegari**

July 2013