பன்மொழி செய்தித் திரட்டுதலை நோக்கி
(Towards Multilingual News Aggregation)

Venkatesh R
Sify Limited,
Tidel Park, II Floor 4, Canal Bank Road, Taramani, Chennai 600 113
Email: venkatesh_r@sifycorp.com

Abstract:

Online news is an important component that adds to the utility value of the Internet. Many newspapers have their web editions along with portals that also generate lot of news content for their readers. Aggregating news stories in English is comparatively easy and a straight-forward task. Samachar (http://www.samachar.com), a pioneer in English news aggregation, has been successfully catering to a lot of Non-Resident Indian readers. Extending the same facility to Indian languages is the theme of this paper. It also deals with the problems faced while implementing and the future course of development, in this area.

I. அறிமுகம்

தமிழர்கள் வாழ்வில் செய்தி என்பதற்கு முக்கிய இடமுண்டு. ஒவ்வொரு நாள் காலையும் செய்தித்தாளும் காபியுமாக பலர் தம் நாளைத் தொடங்குவர். பலருக்கு நாட்டு நடப்புகளைத் தெரிந்துகொள்ளாமல் எதுவுமே முடியாது என்ற அளவு செய்தி என்பது அவர்களை ஆக்கிரமித்துக்கொண்டு உள்ளது. இத்தகைய ஆர்வமும் ஈடுபாடும், தொலைதொடர்பு சாதனங்கள் வளரவளர மேன்மேலும் வளர்ந்துகொண்டே வந்திருக்கிறது என்று சொன்னால் பொருத்தமாக இருக்கும்.

வானொலி, தொலைக்காட்சியுடன் இணை ஊடகமாக வளர்ந்துள்ள இணையம், இன்று செய்திகளை வழங்குவதில் முன்னணியில் உள்ளது. பலர், செய்திகளை "உடனடியாக" வழங்கும் ஓர் ஊடகமாக, இணையத்தைப் பார்க்கத் தொடங்கியுள்ளனர் என்பதை மிகப்பெரிய கௌரவமாகக் கொள்ள வேண்டும். செய்திகளுக்காக இனி யாரும் அடுத்த நாள் காலை வரை காத்திருப்பதில்லை. தொடக்கத்தில், இந்தியா பற்றிய செய்திகளை உடனுக்குடன் தெரிந்துகொள்ள விரும்பும் பெரும்பகுதியினர் வெளிநாடுகளில் வாழும் இந்தியர்களாகவே இருந்தனர். இன்று இணையம் பற்றி தமிழக அளவில் ஏற்பட்டுள்ள கவனம், அதன் பயனைப் புரிந்துகொண்டுள்ள தன்மை ஆகியவற்றினால், பல தமிழ்நாட்டுத் தமிழர்களும் செய்திகளைத் தெரிந்துகொள்ள இணையம் நோக்கி முன்னேறி வருகின்றனர்.

இணையம் என்பது ஒருவிதத்தில் அவசரமான ஊடகமும் கூட. இணையத்தில் யாரும் படிப்பதில்லை என்று இணைய ஆய்வாளர்கள் சொல்கிறார்கள். பயனர்கள் சட்டெனப் பார்வையை ஓட்டி (scan) உடனடியாகத் தகவல்களைத் தெரிந்துகொள்ளவே விழைகிறார்கள். மேலும், இணையம் என்பது தகவல்கள், செய்திகள் கொட்டிக் கிடக்கும் பெரிய கிடங்கு. அதில் தமக்கு வேண்டியவற்றை, ஒப்பிட்டுப் பார்த்து, தேவையானவற்றை எடுத்துக்கொள்வது ஒரு விதம். மற்றொரு விதம், ஒரே செய்தியை பல்வேறு பத்திரிகைகள் வழங்கும் பாணி என்ன, சாய்வுகள் என்ன, கூடுதல் தகவல்கள் என்ன என்பதைத் தெரிந்துகொள்ள விழைவோரும் உண்டு. மேலும்,

ஒரே செய்தியின் அடுத்தடுத்த வளர்ச்சிகளைத் தெரிந்துகொள்ளவும் இணையம் இன்று பெரும்பாலும் பயன்படுகிறது.

இங்கேதான் சமாச்சார்.காம் வலைதளத்தின் பணி தொடங்கியது. சமாச்சார் தொடங்கியபோது, இருந்த மற்றொரு சூழல், குறுகிய பாட்டை (Narrow band) இணைப்புகள். அதனால், பல செய்தித்தாள்களை மாற்றி மாற்றி பார்க்கவேண்டியிருந்தது. அதற்குள், இணைப்பு துண்டிக்கப்பட்டுவிடும். அல்லது, ஒரு பக்கம் கீழறிங்கவே நேரம் எடுத்துக்கொள்ளும். அதனால், பெரும்பாலும், ஒன்றிரண்டு செய்தித்தாள்களைப் பார்த்தாலே போதும் என்ற தேக்கம் ஏற்பட்டிருந்த காலகட்டம் அது.

அந்தச் சமயத்தில்தான் சமாச்சார் வலைதளம், அனைத்து செய்தித்தாள்களின் தலைப்புச் செய்திகளையும் ஒரே பக்கத்தில் கொண்டு வந்து சேர்த்தது. அதுவும் அரை மணி நேரத்துக்கு ஒருமுறை செய்திகள் மாற்றம் பெற்றுக்கொண்டே வந்தன. நாளிதழ்களின் செய்திகள் மாற மாற, இங்கே சமாச்சார் வலைதளத்திலும் அவை பிரதிபலிக்கத் தொடங்கின.

இந்த முறையினால் பல ஆதாயங்கள்.

வாசகர்களுக்கு:

1. ஒரே பக்கத்தில் ஒரு துறைச் செய்திகளை அடுத்து அடுத்து பார்த்துக்கொள்ளலாம்.
2. ஒரே செய்தியின் பல படிவங்களை, வழங்குமுறைகளைப் பார்த்துக்கொள்ளலாம்.
3. ஒரே செய்தியின் வளர்ச்சிகளை?? தொடர்ந்து கவனிக்கலாம்.

நாளிதழ்களுக்கு

1. செய்திகள் மிகப்பெரும் வாசகர்களைச் சென்று சேர்ந்தது.
2. செய்திகளின் தன்மையைப் பொறுத்து, பயனர்களின் ஆர்வத்தைப் பொறுத்து, அவர்கள் வேண்டிய செய்திக்கு நேரடியாகவே வந்து சேர்ந்தார்கள்.
3. அதனால், இணையத்தில் நாளிதழ்களின் பயன்பாடு மிகப்பெரும் அளவில் உயர்ந்தது.

இதுதான் சமாச்சாரின் வெற்றிக் கதை.

இதை மற்ற இந்திய மொழிகளில் செய்து, அதே அளவு முன்னேற்றத்தை உருவாக்கவேண்டும் என்ற ஆர்வமே எங்களை உந்தித் தள்ளியது.

II. இந்திய மொழிகளில் சமாச்சார்

ஆங்கிலத்தில் வெற்றிகரமாக நடைபெற்று வந்த சமாச்சார் வலைதளத்தைப் போன்று ஏன் பிற இந்திய மொழிகளிலும் செய்துபார்க்கக் கூடாது என்று எங்கள் நிறுவனத்திற்குள் யோசனை சுழன்றுகொண்டிருந்தது.

இந்திய மொழிகளில் இதுபோன்ற ஒரு முயற்சி தேவைப்படுமெனில், ஒரே ஒரு முக்கியப் பிரச்சினையைத்தான் நாங்கள் எதிர்கொண்டோம். அது, பல எழுத்துக் குறியீடுகள்.

உதாரணமாகத் தமிழையே எடுத்துக்கொள்வோம். தினமணி, தினமலர், தினகரன், தினத்தந்தி ஆகிய முக்கிய நாளிதழ்ளின் வலைத்தளங்களைப் பார்க்க உங்களுக்கு ஒவ்வொரு எழுத்துரு

தேவை. இயங்கும் எழுத்துரு (Dynamic Fonts) தொழில்நுட்பம் வந்தபின், பலர் தங்கள் வலைளதங்களை அப்படி மாற்றினார்கள். அப்படியும் பல உலாவிகளில், பக்கம் முழுமையாக கீழறிங்கி, சரியான எழுத்துருவில் தெரியாமல் போகும் அபாயங்கள் உண்டு.

தமிழைப் பொருத்தவரை, இணையப் பயன்பாடு பெரும் பாய்ச்சலோடு வளராமல் போனதற்கு அடிப்படைக் காரணங்களில் ஒன்று இந்த எழுத்துருப் பிரச்சினையே. பல அலுவலக கணினிகளில் எழுத்துருவை இறக்கிக்கொள்ள அனுமதி கிடையாது. பல வீடுகளில், எழுத்துருவை இறக்கிக்கொண்டே படிக்கவேண்டியிருக்கும் என்ற எண்ணமே மருட்சியை ஏற்படுத்தி விடுகிறது. அல்லது அப்படி ஒன்று இருக்கிறது என்பதே தெரியாமல் இருக்கிறார்கள். இதில் பயனர்களைக் குறை சொல்லிப் பிழையில்லை. அதேபோல், பல இணைய உலாவி மையங்களிலும் (Cyber cafe's) எழுத்துருவை இறக்கிக்கொள்ள அனுமதி கிடையாது.

இது தமிழுக்கு மட்டும் உண்டான பிரச்சினையாகத் தோன்றவில்லை. இந்திய மொழிகள் பலவற்றில் இத்தகைய பிரச்சினை உண்டு. தனிஆர்வலர்களே ஒவ்வொரு மொழியிலும், தொழில்நுட்ப முயற்சிகளை முன்னெடுத்துச் சென்றிருக்கிறார்கள். அவர்களுக்கு அன்றைய தேதியில் தெரிந்த தொழில்நுட்பத்தைக் கொண்டு, உரிய முறையில், தமது வலைதளத்தைக் கட்டமைத்துக்கொடுத்திருக்கிறார்கள்.

நாளிதழ்கள் காலில் சுடுநீரைக் கொட்டிக்கொண்டு ஓடுகின்றன. அவர்கள் ஒரே எழுத்துரு, ஒரே குறியீடு போன்ற செய்திகளில் கவனம் செலுத்துவார்கள் என்று எண்ணம் எனக்கில்லை. மேலும், நண்பர்கள் பலரோடு பேசும்போது கிடைக்கும் ஒரு தகவல், விற்பனைக்குப் பின்னான உதவி. இன்று தமிழகத்தில் எழுத்துருவையும் செயலிகளையும் விற்பனை செய்பவர்கள், தொடர்ந்து இந்த உதவிகளை வழங்கி வருகிறார்கள். அல்லது, இப்படிச் சொல்லலாம். அப்படி விற்பனை செய்பவர்களே, முழு வலைதளத்தையும் அமைத்துத் தர உதவுகிறார்கள். அப்படி இருக்கும்போது, அவர்கள் ஒரே குறியீடு போன்ற செய்திகளில் கவனம் செலுத்துவது போன்று தெரியவில்லை.

இதுதான் நிஜம். இதை ஏற்றுக்கொண்டு மேலே செல்வது எப்படி என்றுதான் நாங்கள் யோசித்தோம். அப்போதுதான், ஒரு குறியீட்டில் இருந்து மற்றொன்றுக்கு உடனடியாக மாற்றும் வசதி உண்டு என்று தெரியவந்தது. அப்படியிருக்குமானால், இந்தப் பணியைச் சுலபமாகச் செய்யலாம் போலிருக்கிறதே என்று ஆர்வம் எங்களை மேலும் இந்த முயற்சியில் இறங்க வைத்தது. அன்றைய தேதியில், இன்று பேசுவதுபோன்ற ஒருங்குறியான யூனிகோட் எழுத்துருக்கள், அவ்வளவாகத் தெரியவரவில்லை. மேலும், சின்னதொரு குழப்பமும் நிலவிவந்த காரணத்தால், அனைத்து எழுத்துருக்களையும் சிஃபி பயன்படுத்தும் எழுத்துருவுக்கே மாற்றிவிடுவது சுலபமான வழியாகத் தோன்றியது.

III. சமாச்சார் எப்படிச் செயல்படுகிறது

இந்தப் பணியில் எங்களுக்குத் தொழில்நுட்ப ரீதியாக உதவியவர்கள் அண்ணா பல்கலைக்கழகம், கே.பி.சந்திரசேகரன் சென்டர் (AU-KBC).

இந்திய மொழிகளில் எப்படி சமாச்சார் வேலை செய்கிறது என்பதைப் பார்ப்போம். உதாரணமாகத் தமிழை எடுத்துக்கொள்வோம். பின்வரும் வகையில் இந்த சேகரிப்பு நடைபெறுகிறது

1. அரை மணி நேரத்துக்கு ஒரு முறை, தேடு (crawler) பொறி, கொடுக்கப்பட்டுள்ள சுட்டி - தினசரிகளின் செய்திப் பக்கத்தின் சுட்டி - போய்ப் பார்த்து, அந்த HTMLஐக் கொண்டு வரும்.
2. பின் அதில் உள்ள HTML CODEகளை எல்லாம் நீக்கிவிடும்.
3. பின் அதில் உள்ள தலைப்புச் செய்திகளை இனம் காணும்.
4. முதலில் உள்ள 5 தலைப்புச் செய்தி வரிகளை மட்டும் எடுத்துக்கொள்ளும்.
5. பின், ஏற்கனவே குறிப்பிட்டுள்ளபடி, அந்த எழுத்துருவின் குறியீட்டில் இருந்து, சமாச்சார் தமிழ் பயன்படுத்தும் குறியீட்டுக்கு உடனடியாக மாற்றிவிடும்.
6. அதை அப்படியே மற்றொரு HTMLஆக உருவாக்கி, சமாச்சார் தமிழ் வலைதளத்தில் கொண்டு வந்து காண்பிக்கும்.
7. கடைசியாகக் கட்டமைக்கப்படும் HTML மட்டும் சேமிக்கப்படும். ஏழுநாள் களஞ்சியத்தை உருவாக்க, இந்தக் கடைசியாக உருவாகும் பக்கத்தை மட்டும் வைத்துக்கொள்ளும்.

IV. எதிர்கொண்ட பிரச்சினைகள் / தீர்வுகள்

எதிர்கொண்ட- எதிர்கொள்ளும் தொழில்நுட்பப் பிரச்சினைகள்

1. முதல் பிரச்சினை, இன்றுவரையும் நீடிக்கும் பிரச்சினை என்றால் அது ஒன்றுதான். சமாச்சார் தமிழ் வலைதளத்தில் வந்து படிக்கிறவர்களுக்கு, தலைப்புச் செய்திகளை மட்டும் படிக்க முடியும். அதுவும் ஒவ்வொரு அரைமணி நேரமும் மாறிக்கொண்டே இருக்கும். தேவைப்படும் ஒரு செய்தியின் சுட்டியைத் தட்டி, நாளிதழின் வலைதளத்துக்குப் போகும்போது, அங்கே இயங்கும் எழுத்துரு இருக்குமானால், சுலபமாகப் படிக்க முடியும். அது இல்லையென்றால், மீண்டும், எழுத்துரு பிரச்சினைதான்.

இதற்கான தீர்வாக இப்போதைக்கு, நாங்கள் அனைத்து நாளிதழ்களின் எழுத்துருக்களையும் ஒரே கோப்பில் சேமித்து வைத்திருக்கிறோம். மொத்தமாக அந்த எழுத்துருக்களை?? கீழிறக்கிக்கொள்ளலாம்.

அனைத்து நாளிதழ் தளங்களும் ஒருங்குறிக்கு மாறும்வரை, இந்தப் பிரச்சினை இருக்கவே செய்யும்.

2. சில நாளிதழ்களின் வலைதளங்கள் குறித்த நேரத்தில் அப்டேட் செய்யப்படாமல் போய்விட, சமாச்சார் தமிழ் வலைதளத்தில் அவை எந்த செய்தியையும் காட்டாமல் வெறுமையாக இருக்கும்.

தீர்வு: ஒவ்வொரு முறையும் தேடுபொறி (crawler) ஓடி, பெறப்படும் தினசரிகளின் வலைதளத்தில் ஏதும் புதுச் செய்திகள் இல்லையென்றால், கடைசி அரைமணி நேரத்தில் சேகரிக்கப்பட்ட செய்தியையே வைத்துக்கொள்வது என்று தீர்மானிக்கப்பட்டது.

ஏனெனில், பல நாளிதழ்கள், ஒரு நாளைக்கு ஒருமுறை மட்டுமே அப்டேட் செய்யும் வழக்கத்தைக் கொண்டுள்ளன. அவர்கள் தங்கள் நாளிதழ் தயாரிப்பை இரவு ஒன்று, ஒன்றரை மணிக்கு முடித்துவிடுகின்றன. அத்தோடு, நேரடியாக நாளிதழ்கள் அச்சுக்குப் போய்விட, மற்றொருபுறம், வலைதளத்தில் அச்செய்திகள் வலையேற்றப்பட்டு விடுகின்றன. மீண்டும் அடுத்த நாள்தான் செய்திகள் புதுக்கம் பெறும்.

3. சில தினசரிகள் ஏதேனும் தொழில்நுட்பக் காரணங்களால், செய்திகளை அப்டேட் செய்யாமல் நிறுத்தி வைத்துவிடுதல்.

தீர்வு: இந்தப் பிரச்சினை பல சமயங்களில் ஏற்படுவதுண்டு. குறிப்பாக, முகப்புப் பக்கத்தில் உள்ள அரசியல், தேசிய, மாநில செய்திகள் தப்பித்துவிடும். நகர பதிப்புச் செய்திகளில் இந்தப் பிரச்சினை தலைகட்டும். அங்கே, பழைய செய்திகளைக் காண்பிப்பது சரியாக இராது என்பதால், முதலில் SITE NOT UPDATED என்றொரு குறிப்பை வழங்கி வந்தோம். பின்னர் இதுவும் சரியாக இராததால், அப்டேட் ஆகாத நாளிதழ்களின் குறிப்பிட்ட இணைப்புகளை அப்போதைக்கு நீக்கி விடுகிறோம். பின்னர், நாளிதழின் வலைதளம் இயங்க ஆரம்பித்த பின், மீண்டும் இணைத்துக் கொள்கிறோம்.

4. மற்றொரு முக்கியப் பிரச்சினை: தினசரிகள் பல சமயங்களில் தமது டெம்பிளேட்டுகளை மாற்றிவிடுதல்.

தீர்வு: முதலில் பெரியப் பிரச்சினையாக உருவானது இதுதான். ஏதோ காரணங்களால், தாங்கள் செய்தி வழங்கும் இடைமுகத்தை, நாளிதழ்கள் மாற்றி வருகின்றன. அதைப் புரிந்துகொள்ளும் HTML DECODERஐ உருவாக்கினோம். இது அனைத்து இடைமுக மாற்றத்தையும் சமாளிக்கும் திறனுடையதாக்கினோம். அதன் மூலம், பெறப்படும் பக்கத்தில் (TARGET SITE) எவ்வகையான மாற்றம் ஏற்பட்டாலும், அதையும் உள்வாங்கிக்கொண்டு, தலைப்புச் செய்தி வரிகளை மட்டும் பெறும் வசதி உருவானது.

5. மற்றொரு பிரச்சினை: பக்கங்களுக்குப் பெயரிடும் முறையை மாற்றுவது (Changing the naming convention)

உதாரணமாக, இப்படி ஒரு சுட்டி

http://www.dinakaran.com/daily/2004/Sep/13/index.html

பார்த்தாலே உங்களுக்குப் புரியும். Sep என்று இச்சுட்டியில் உபயோகிக்கப்படும் கோப்பின் பெயர், அடுத்த முறை, அதே மூன்று எழுத்துக்களோடு, அதாவது Oct என்றிருந்தால், பிரச்சினை இல்லை. அதுவே Octo போன்ற வகையில் பயன்படுத்தப்படுமாயின், மீண்டும் தேடுபொறி (crawler) இயங்காது. இதையும் உள்ளடக்கி, பொதுவான தேடும்திறனுள்ள பொறியை உருவாக்கினோம்.

6. எழுத்துருவையே மாற்றுவது

சில நாளிதழ்கள், தமது எழுத்துருக்களை மாற்றி வருகின்றன. தாம் குறியீட்டின் பல வடிவங்களில் உள்ள எழுத்துருக்களுக்கே இவை மாறியிருக்கின்றன. அப்படி மாறும்போது, வேறு வழியே கிடையாது. ஏற்கனவே உள்ள குறியீடு என்றால் கொஞ்சம் வேலை சுலபம். இல்லை புது குறியீடு என்றால், அதற்காக கன்வர்டர்களை உருவாக்கியே ஆகவேண்டும்.

இதெல்லாம் பிரச்சினைகள் என்று சொல்வதை விட, நல்ல பாடங்கள். நிறைய கற்றுக்கொண்டோம்.

அப்படிக் கற்றுக்கொண்டதன் தொடர்ச்சிதான், இதே போன்ற செயலிகளைப் பிற இந்திய மொழிகளுக்கும் செய்யத் தூண்டியது. முக்கியமாக முதலில் இந்தி, தெலுங்கு, மலையாளம், கன்னடம் ஆகிய மொழிகளை எடுத்துக்கொண்டு, அந்தந்த மொழிகளில் உள்ள முக்கிய

தினசரிகளின் சுட்டிகளைச் சேகரித்து, அவற்றுக்கான எழுத்துருக்கள், அதற்கான கன்வர்ட்டர்கள் ஆகியவற்றையும் உருவாக்கினோம்.

பிற மொழிகளில், தெலுங்கு மொழியில் மட்டுமே பெரிய பிரச்சினையை எதிர்கொண்டோம். எண்ணற்ற எழுத்துக்களையும் துணை குறிகளையும் கொண்ட மொழி தெலுங்கு. ஒரு எழுத்துருவில் உள்ள வடிவம் மற்றொன்றில் வேறொரு மாதிரியிருக்கும் அதிசயத்தையும் அங்கேதான் கண்டோம். அதையும் புரிந்துகொண்டு விரைவில் கன்வர்டர்களை உருவாக்கினோம்.

மலையாளம் மற்றும் கன்னடம் ஆகிய மொழிகளில் சமாச்சாரை உருவாக்கும்போது கண்ட மற்றொரு முக்கியப் பிரச்சினை, பெறப்படும் நாளிதழ் தளத்தின் ஒவ்வொரு இணைய பக்கமும் ஒவ்வொரு டெம்பிளேட், இடைமுகத்தைக் கொண்டிருப்பது. அதையும் உள்ளடக்கி, செய்திகள் திரட்டுவதை செம்மைப்படுத்தியிருக்கிறோம்.

V. அடுத்த கட்ட வளர்ச்சி

இரண்டு முன்னேற்றங்களை எதிர்நோக்குகிறோம்.

1. யூனிகோட் அனைவராலும் ஏற்கப்பட்டு வரும் இவ்வேளையில், அனைத்துச் செய்திகளையும் யூனிகோட் எழுத்துருவில்
வழங்க முற்படுவது. அதன் மூலம் தேடுவது சுபலமாகும்.

இன்றைக்கு, இந்திய மொழிகளில் உள்ள குழப்பத்துக்குக் கிடைத்திருக்கும் ஒரே தீர்வு ஒருங்குறிதான். ஒருங்குறியும் இயங்கும் எழுத்துருவும் இணைந்து, படிக்கும் வசதியை மேம்படுத்தும்.

இதில் உள்ள ஆதிப் பிரச்சினை அப்படியேதான் இருக்கின்றன. செய்திகளைப் பெறும் நாளிதழ் வலைதளங்கள் தங்களை ஒருங்குறிக்கு விரைவில் மாற்றிக்கொள்ளாவிட்டால், மீண்டும், அதே எழுத்துரு பிரச்சினை நீடிக்கவே செய்யும். விரைவில், நாளிதழ் வலைதளங்கள் மாறுவதே புதிய பாய்ச்சலுக்கு வலு சேர்க்கும்.

2. அடுத்து, பன்மொழி செய்தி திரட்டுதல். மொழி எல்லைகளைக் கடந்து, அனைத்து மொழி நாளிதழ்களின் தலைப்புச் செய்திகளையும் ஒரே பக்கத்தில் கொணர்வது. ஒரு கட்டத்தில், நமது சேகரிப்பு செயலி மேலும் புத்திசாலித்தனமாகுமென்றால், ஆங்கிலச் செய்திக்கு இணையான பிற இந்திய மொழிச் செய்திகளையும் பக்கத்தில் வைக்க முடியும். இயந்திர மொழிபெயர்ப்பு (Machine Translation) சாத்தியப்படும்போது, இதற்கு மேலும் ஒரு பரிமாணம் கிடைக்கும்.

****